Journal of Applied Psychology

Edited by

Donald G. Paterson
University of Minnesota

Volume 33, 1949

Contents of Volume 33

Articles

iii

Book Reviews

Miscellaneous

Attitude Research in the Army *

Jack Elinson

*Troop Attitude Research Branch, Troop Information and Education
Division, Special Staff, United States Army*

Many Army policies affecting troops depend on soldier reactions and cooperation for success. Necessarily then, those formulating policies need to know soldier reactions. Obviously, the larger an organization is, the harder it is for top management to keep in touch with what the troops (or employees) are thinking. Attitude surveys and opinion polls of soldiers are a carefully developed means of helping higher headquarters keep well-informed on these matters—as well informed as a good company commander or good supervisor can be as a result of getting around in his company or unit and talking and working with his men. Such surveys can determine:

In case of an existing policy . . .
> . do men know about it and understand it?
> . are they in sympathy with it?
> . do they feel it is being carried out as intended?

In case of a proposed policy . . .
> . what is likely to be its effect?
> . how are men likely to react to it?

Research on troop opinion helps provide answers to such questions.

Broadly speaking, attitude research functions for Army administration in the following five ways:

1. As a means of anticipating troop reaction to a new administrative policy.

---

* This article was originally prepared as an administrative memorandum for the use of Lieutenant General Willard S. Paul, Director, Personnel and Administrative General Staff, USA. As such it represents the thinking of Major Paul D. Guernsey, Chief, Troop Attitude Research Branch, and Ira H. Cisin, Sr. Analyst in Charge of Unit Studies, as well as that of the writer, who is Sr. Analyst in Charge of Surveys.

2. As a guide in the formulation of administrative policy or of a change in policy.

3. As a means of evaluating the operation of an existing administrative program.

4. As a means of evaluating, experimentally, the effectiveness of an information or training program.

5. As a source of quantitative information and evidence in support of or against a proposed policy or change in policy.

Anticipating Troop Reaction

1. *Troop Attitude Toward Army's New Career Guidance Program.* The Army's new Career Guidance Program for enlisted personnel involves the establishment of a systematic promotion ladder within each type of service. Advancement up the ladder will be based essentially on Army-wide competitive testing. It was planned that the program would first go into effect for men in the Infantry. Before all the details of the Career Guidance Program had been decided upon, an attitude study was conducted among Infantrymen in order to get a preview of their reactions to the plan.

The study indicated that: although the new Career Program was acceptable to enlisted men *in principle*, many enlisted men were in opposition to some of the details of the proposed program. In addition to revealing attitudes of men toward the various phases of the Career Program, the survey also disclosed areas of ignorance about the Career Program. So that, while attitudes toward the Career Program may be difficult to change and some alteration in the administrative details of the program may appear necessary, areas of ignorance about the program may be skillfully attacked with well-directed informational activity.

Formulation of Administrative Policy

2. *Troop Attitude on Order of Demobilization.* Months in advance of VE-Day, the War Department's Special Planning Division was anticipating the likelihood that demobilization policy adopted on defeat of Germany could result in morale disaster if the plan adopted were to be far out of line with what troops would consider fair.

The problem then was to determine accurately what plan troops would be likely to consider fair. Troop cross sections were surveyed by research teams in the United States and in overseas theaters as early as November 1943 and several times subsequently.

Research revealed four factors to be critical in soldiers' minds: (1) length of service; (2) time overseas; (3) parenthood; and (4) combat participation.

These were the four basic factors adopted in the Adjusted Service Rating Plan (Point Score for demobilization).

One can read today in the book, just published, written by the Historical Division, Department of the Army, entitled *The Army Ground Forces in World War II*, how considerations of military necessity as well as troop attitudes dovetailed into final determination of administrative policy.

Evaluating an Existing Program

3a. *Trend Surveys in the Universal Military Training Experimental Unit.* When the Universal Military Training Experimental Unit was set up at Fort Knox, Kentucky, the program included various innovations in military procedure: Code of Conduct (a form of demerit system), Trainee Courts, with men themselves sitting somewhat as a jury, considerable emphasis on the Chaplains' activities, compulsory educational program, concerted attention to off-duty activities, etc. In order to measure the trend of trainee reaction to these innovations, and to the training program in general, attitude studies have been made among the trainees in each of the cycles going through the unit,—studies conducted at the beginning, the middle and the end of each training cycle. The attitudes of the officers and cadre of the unit have also been obtained at the end of each training cycle. From the reports on these studies the Commanding General of Army Ground Forces and the Commanding General of the unit have followed any shift in reactions as trainees progressed through their training. As modifications are made in the experimental training program at the unit, the studies are re-designed to evaluate the results.

3b. *Studies Pertaining to Recruitment for the Military Service.* In the Fall of 1947, staff officers of the Army's Military Personnel Procurement Service Division and their advertising agency, the N. W. Ayer Co., began to feel as a result of a continuing decline in enlistments that a change in advertising direction was indicated.

Accordingly, two coordinated surveys were conducted: the first by the Army, through its Attitude Research Branch to survey newly enlisted recruits; the other by the advertising agency through a commercial polling organization to survey young civilian males and their parents.

The surveys yielded new insights into the problem of appropriate advertising for recruiting. For example, one traditional advantage of military service—early retirement and good retirement pay—was found to have practically no appeal among 17–18 year old youngsters, but was of considerable importance in the re-enlistment of older veterans.

Evaluating an Information Program

4. Most staff sections need at one time or other to have troops in the field informed on certain matters. Questions looming in the minds of those who must get out information are: (a) how can the information be made to reach the largest proportion of those who should be reached? (b) what presentation will be most effective in getting the information read after it is gotten out? (c) how can the information be put across so that it is most likely to be remembered once it is read? and once the information is released, (*1*) how widely has it been seen, read and remembered? (*2*) did it accomplish what it was supposed to accomplish?

Effectiveness of any single information tool or device, such as movies, radio programs, posters, pamphlets, training courses, and the like, can validly be determined only by a true experimental approach using as subjects both control and experimental groups. During the war, numerous such studies were made by the Research Branch on Hollywood-produced films which were calculated to give the soldier a better understanding of the issues of the war. Compared to broad cross-sectional sample surveys, experimental evaluation studies of this kind are usually less costly, but they remain inordinately extravagant in the use of research personnel time, and also involve more than the usual cooperation of operating officials, that is, commanding officers. Consequently, since the war, such studies of information media have been restricted to those of exceptional importance. One, currently under way, is an experimental evaluation of the new film produced under the auspices of the Surgeon General of the Army, entitled "Miracle of Living," a film designed to produce certain changes in information, attitude, and behavior among enlisted men with respect to venereal disease.

Evidence Pro and Con of a Proposed Policy or Change in Policy

5. Virtually *all* Research Branch studies have been used or are potentially useful for the purpose of providing a source of quantitative information and evidence in support of or against a proposed policy or a change in policy. In contrast to arm-chair opinion based on umbilical meditations, quantitative evidence derived from scientific sampling surveys are invaluable tools in the hands of skillful administrators. Among instances which may be mentioned of this use of attitude research data are studies of officer-enlisted man relationships used by the Doolittle Board in preparing its recommendations, attitudes toward Army Courts-martial procedure, survey of educational and recreational interests of soldiers, surveys among hospital patients with respect to treatment, surveys for the Quartermaster General on soldiers' food and clothing preferences,

comparison of competing physical training programs, survey among medical officers with respect to reasons why they would or would not accept commissions in the Regular Army, housing demands both among men in the Army and those about to be discharged, attitudes of officers toward logistical careers and training programs, and other studies of a more confidential nature. In short, as General Lanham has phrased it, attitude research has, within small and useful margins of error, proved itself to be the "morale radar" of the Armed Forces.

The Psychological Barometer on Communism, Americanism and Socialism

Henry C. Link and Albert D. Freiberg

The Psychological Corporation, New York City

The following results are taken from three Barometer surveys: the August, 1948 survey made with 5000 urban interviews; the October, 1948 survey made with 1000 interviews but with a comparable sample; the November, 1948 survey made with 10,000 urban interviews. The dates and size of sample are given with each table.

In the August Barometer of 5000 interviews, one of the questions asked was:

Q. What, in your opinion, are the three most dangerous threats within our own country to a prosperous America?

This question was asked specifically for the employee relations division of the General Electric Company. The answers showed that two threats, inflation and Communism, were considered by far the most dangerous, with strikes and industrial conflict a distant third. The per cents mentioning various dangers were:

Threats to a Prosperous America

Answers, August 1948	%
Inflation, high prices	49.5
Other economic threats such as a depression, 4.4%; high or low wages, 1.7%; O.P.A. or lack of O.P.A., .9%; miscellaneous, 1.8%; total	8.8
Communism	44.1
Fascism, .9%; Socialism, .5%; foreign spies and infiltration, 1.1%; lack of freedom, .3%; total	2.8
Strikes, struggle between capital and labor	12.1
Power of unions, organized labor	5.0
Taft-Hartley Act	.3
Politicians, political parties, politics	10.6
War talk, threat of war	10.6
Big business, monopolies, Wall Street, capitalism, high profits	2.9
Race prejudice and intolerance	8.8
Civil rights program, Jews, Negroes, immigrants	1.7
Atomic bomb, 1.8%; inadequate military defense, .6%; draft, .4%; poor foreign policy, E.R.P., 1.4%; the Russians, 1.6%; total	5.8

9

Social and psychological threats:
 Lack of housing, 4.2%; alcohol, drinking, 3.8%;
 crime, delinquency, 3.6%; lack of religion, 3%;
 family trouble, 1.9%; poor education, 1.9%;
 movies, theatres, radios, comic books, .6%; lack

of cooperation, 3%; greed, 1.7%; misc. 5.8%; total	29.5
Bad govt., bureaucracy, graft, govt. racketeers, govt. restrictions, govt. spending, high taxes; total	7.5
Natural disasters including fire, floods, rodents, drought, wastefulness of resources	3.5
Miscellaneous	11.3
Don't know	12.7
Total Interviews	5000

Is Communism Becoming Dangerous?

The growing danger of Communism in the United States is further indicated by the answers in 1946 and in 1948 to this question:

Q. It is being said that Communism is becoming a dangerous thing in the United States. Do you think this is true or not?

Answers	April 1946	October 1948
	%	%
True	51.2	67.0
Not true	34.1	24.5
Don't know	14.7	8.5
Total Interviews	2500	1000

This conviction is shared pretty much by all socio-economic groups, and by union and non-union families alike, as shown by the following table:

	Socio-Economic Group				Union Membership	
Answers, Oct. 1948	A	B	C	D	Union	Non-Union
	%	%	%	%	%	%
True	80	66	67	62	65	68
Not true	14	26	27	24	29	23
Don't know	6	8	6	14	6	9
Total Interviews	100	300	400	200	278	722

Are Communists Traitors?

In previous surveys, it was found that Communists in the United States were regarded by 77 per cent to be a fifth column, loyal to Russia first, rather than as a typical American political party. A majority favored outlawing the Communist party. The sharpest definition of this issue was made in the question:

Q. Do you think a Communist is a traitor to the United States?

Answers	January 1948	October 1948
	%	%
Yes	65	70.6
No	18	18.0
Don't know	17	11.4
Total Interviews	000	2500

Union and non-union members thought alike on this subject, whereas, by socio-economic groups, the "yes" answers ranged from 81 per cent in the "A" group to 62 per cent in the "D" group.

Is Socialism Becoming Dangerous?

Whereas Communism in the last two years has been sharply recognized by the American people as a threat to their institutions, their reactions to Socialism are quite different. Where 67 per cent say that Communism is becoming dangerous, only 26 per cent say that Socialism is becoming dangerous.

Q. It is being said that Socialism is becoming a dangerous thing in the United States. Do you think this is true or not?

Answers, Oct. 1948	Total	Socio-Economic Group				Union Membership	
		A	B	C	D	Union	Non-Union
	%	%	%	%	%	%	%
True	26.4	30	34	23	20	21	28
Not true	50.5	54	50	54	43	53	50
Don't know	23.1	16	16	23	37	26	22
Total Interviews	1000	100	300	400	200	278	722

Are Communism and Socialism the Same?

Because of these widely different reactions toward Communism and Socialism, this further question was asked:

Q. Do you think that Socialism and Communism are about the same or are they different?

Answers, Oct. 1948	Total	Socio-Economic Group				Union Membership	
		A	B	C	D	Union	Non-Union
	%	%	%	%	%	%	%
Same	22.8	16	21	26	22	28	21
Different	60.9	73	66	59	53	57	62
Don't know	16.3	11	13	15	25	15	17
Total Interviews	1000	100	300	400	200	278	722

Union members are more likely to regard them as the same than are non-union members, but the higher the educational level, the more likely people are to regard them as different. In answer to the question:

Q. What difference do you think there is between them?

Some of the principal reasons given were: Communism is totalitarian while Socialism isn't; Socialism recognizes individual rights; Socialism is more liberal, more democratic; Communism means force, Socialism does not; Socialism is gradual, Communism is revolutionary; Communism is bad, Socialism is good, etc., etc. However, 39 per cent gave no answer.

Specific Issues on Communism and Socialism

The sharp repudiation of Communism as compared with Socialism is no doubt influenced by the strained relations between Russia and the United States. Therefore, it is of unusual significance to ascertain people's reactions to specific measures which tend to bring about Socialism or Communism, or both, in this country. In the previous survey,[1] we reported on such issues as government versus private ownership of manufacturing companies, who does the most for the good of the workers, preference for jobs in private industry or the government, and investing money in government bonds or private concerns.

One of the questions asked in the October survey was:

Q. Do you think government control of business would be a step toward Communism? Toward Socialism?

Answers, Oct. 1948	Toward Communism	Toward Socialism
	%	%
Yes	61.3	49.7
No	22.5	18.9
Don't know	16.2	31.4
Total Interviews	1000	1000

The answers by union membership and socio-economic group to these two questions were:

Q. Do you think government control of business would be a step toward Communism?

Answers, Oct. 1948	Socio-Economic Group				Union Membership	
	A	B	C	D	Union	Non-Union
	%	%	%	%	%	%
Yes	64	65	63	50	59	62
No	21	23	23	23	23	23
Don't know	15	12	14	27	18	15
Total Interviews	100	300	400	200	278	722

[1] Link, H. C. and Freiberg, A. D. The 97th psychological barometer. *Journal of Applied Psychology*, 1948, 32, 443–451.

Q. Do you think government control of business would be a step toward Socialism?

| Answers, Oct. 1948 | Socio-Economic Group | | | | Union Membership | |
	A	B	C	D	Union	Non-Union
	%	%	%	%	%	%
Yes	63	58	48	35	41	53
No	16	19	19	20	21	18
Don't know	21	23	33	45	38	29
Total Interviews	100	300	400	200	278	722

Not inconsistent with the answers to the question on the differences between Communism and Socialism were the answers to the following question:

Q. Do you think a country can have democracy without having private capitalism?

| Answers, Oct. 1948 | Total | Socio-Economic Group | | | | Union Membership | |
		A	B	C	D	Union	Non-Union
	%	%	%	%	%	%	%
Yes	20.9	22	19	21	24	24	20
No	57.4	61	65	58	42	50	60
Don't know	21.7	17	16	21	34	26	20
Total Interviews	1000	100	300	400	200	278	722

More than 42 per cent are either uncertain or say that private capitalism is not necessary for democracy. This is especially interesting in view of the recent statements by Dwight D. Eisenhower, in his installation address as President of Columbia University and other talks, to the effect that private property rights in the United States are the keystone of all other democratic freedoms.

Price Control and the O.P.A.

The readiness of the people to accept socialistic controls, or governmental controls which amount to the confiscation of property, is illustrated by the answers to this question:

Q. What do you think would do most to keep prices down: the O.P.A. and its price ceilings, or free competition by business without any O.P.A.?

| Answers, November 1948 | Total | Socio-Economic Group | | | | Union Membership | |
		A	B	C	D	Union	Non-Union
	%	%	%	%	%	%	%
O.P.A. and its price ceilings	41.5	33	35	43	52	49	38
Competition by business without any O.P.A.	44.5	58	52	43	30	37	48
Don't know	14.0	9	13	14	18	14	14
Total Interviews	5000	500	1500	2000	1000	1438	3562

The opinions of people on price control have been subject to very wide fluctuations. In the spring of 1946, all polls showed a large majority of the public favoring the O.P.A. By the fall of 1946, this attitude had almost completely reversed itself. The results of our polls on this subject are:

Answers	Oct. 1946	Aug. 1948	Nov. 1948
	%	%	%
O.P.A. and its price ceilings	26.1	47.2	41.5
Competition by business without any O.P.A.	65.1	39.7	44.5
Don't know	8.8	13.1	14.0
Total Interviews	2500	5000	5000

Socialistic Trends in Housing

A further illustration of people's readiness to accept socialistic measures is provided by their answers to this question on housing:

Q. How do you think the housing problem will be settled best: (a) by having the Federal Government furnish the money and plans, or (b) by leaving it to private individuals and builders?

Answers, Oct. 1948	Total	Socio-Economic Group				Union Membership	
		A	B	C	D	Union	Non-Union
	%	%	%	%	%	%	%
Having Federal Govt. furnish money and plans	37.0	28	30	39	48	44	34
Leaving it to private builders and individuals	51.8	64	59	51	36	44	55
Don't know	11.2	8	11	10	16	12	11
Total Interviews	1000	100	300	400	200	278	722

Other issues bearing on the conflict between Communism, Socialism, and traditional Americanism or a democracy based on private capitalism will be taken up from time to time.

Attitude Toward the Taft-Hartley Law

The feeling against the Taft-Hartley law among union members or union families is not nearly as unanimous as union leaders have presented it to be. Of those questioned, 94 per cent answered "yes" to the question: Have you heard of the Taft-Hartley law which was passed by Congress to regulate unions, control strikes and get rid of Communist leaders? Then we asked:

Q. During the past year do you think this law has done more harm than good or more good than harm?

| Answers, Oct. 1948 | Total | Socio-Economic Group | | | | Union Membership | |
		A	B	C	D	Union	Non-Union
	%	%	%	%	%	%	%
More harm than good	24.8	15	23	20	24	34	21
More good than harm	39.7	60	50	34	25	29	44
Don't know	35.5	25	27	37	51	37	35
Total Interviews	1000	100	300	400	200	278	722

The Chief Victims of the Increase in the Cost of Living

The answers to this question show one of the sharpest differences by socio-economic groups that we have ever recorded.

Q. Who has suffered most from the increase in the cost of living: the workers on salaries and wages, or the people who must live on the income from life insurance, Government bonds, stocks and other savings?

| Answers, Oct., 1948 | Total | Socio-Economic Group | | | |
		A	B	C	D
	%	%	%	%	%
Workers on salaries and wages	36.3	25	27	38	52
People who must live on income from life insurance, etc.	56.8	69	69	57	33
Don't know	6.9	6	4	5	15
Total Interviews	1000	100	300	400	200

Family Prosperity

Q. Is your family more prosperous (or better off) today than two years ago, less prosperous, or the same?

In spite of high prices, most families continue to think of themselves as better off or as well off as they were two years ago.

| Answers, November, 1948 | Total | Socio-Economic Group | | | | Union Membership | |
		A	B	C	D	Union	Non-Union
	%	%	%	%	%	%	%
More prosperous	24.2	23	25	23	25	25	24
The same	45.8	49	46	40	44	43	47
Less prosperous	26.0	25	25	27	27	29	25
Uncertain	4.0	3	4	4	4	3	4
Total Interviews	5000	500	1500	2000	1000	1438	3562

The above figures show a rather significant difference between the opinions of union members and non-union members. Although the

unions are organized to obtain quick and broad wage increases, union members do not consider themselves as well off in the scale of living as do non-union members who have had to rely on themselves. Contrary to the popular belief that the white collar workers are the principal losers from the cost of living rise, this group, principally the "B" group, considers itself better off than does the large group of skilled and semi-skilled wage workers where unionism is strongest (groups "C" and "D"). This may be due in part to the steadiness of their work as compared with the time lost by wage earners through strikes, material shortages and the indirect results of strikes in related industries.

We have now been asking this question for several years and some of the results are as follows:

Answers	Oct. 1944	Oct. 1943	Oct. 1945	Apr. 1946	Oct. 1946	Apr. 1947	Oct. 1947	Nov. 1948
	%	%	%	%	%	%	%	%
More prosperous	38	29	32	26	31	29	24	24
The same	47	46	51	48	44	42	46	46
Less prosperous	15	23	15	24	22	26	28	26
Don't know		2	2	2	3	3	2	4
Total Interviews	2000	2500	2500	2500	2500	2500	2500	5000

Probability of Another War

The prospects of avoiding war, in people's opinion, have improved during the past year, as shown by the October, 1948 survey. The question was:

Q. Do you think we can make a lasting peace or do you think that there will be another war within the next 20 years or so?

Answers	Feb. 1943	Oct. 1944	Oct. 1945	Oct. 1946	Oct. 1947	Oct. 1948
	%	%	%	%	%	%
Lasting peace	47	28	28	18	11	20
Another war within 20 years	43	54	59	74	77	69
Don't know	10	18	13	8	12	11
Total Interviews	2500	2500	2500	2500	2500	1000

Another question on this same subject was:

Q. How about the next three or four years: another war or no war?

Answers, October 1948	%
War	35.3
No war	42.8
Don't know	21.9
Total Interviews	1000

The Civil Rights Issue

In view of the great controversy over the civil rights program, the following question was asked with interesting results:

Q. Which would do more good for American Negroes: (a) passing laws to give them equal rights with whites; (b) a program to teach white and Negro to get along together?

Answers, October, 1948	Total	Socio-Economic Group				Geographic Area			
		A	B	C	D	East	Mid-West	South	Far West
	%	%	%	%	%	%	%	%	%
Passing laws for equal rights	·11.7	11	12	9	18	12	13	11	9
A program to teach whites and Negroes to get along	77.6	76	80	81	67	77	78	79	77
Neither	1.1	2	*	1	2	1	*	2	2
Don't know	9.6	11	8	9	13	10	9	8	12
Total Interviews	1000	100	300	400	200	370	315	205	110

* Less than .5%

Explanation of the Surveys

Each of these surveys was made with a true cross-section of the urban population. The August and November surveys were made in 100 cities and towns; the October survey was made in 47 cities and towns.

Sampling Methods. A modified area sampling method was used. All interviews were assigned by the local supervising psychologist by blocks and streets in accordance with maps constructed to designate the proper socio-economic levels. These maps are made to divide the population into four principal groups, the "A" group consisting primarily of owners and executives, the "B" group, primarily white-collar and semi-professional, the "C" group or skilled factory and transportation workers, "D" group or the less skilled. All interviews were made in the home, but only one in a family; half were made with women, half with men.

An Analysis of the 1948 Polling Predictions *

Daniel Katz

Survey Research Center, University of Michigan

After naming the winning candidate successfully in three presidential elections, the public opinion polls stumbled badly in 1948 in their unanimous forecast of a Dewey victory. With the exception of the Roper poll, however, the 1948 performance, from an arithmetic point of view, was not as startlingly different from previous forecasts as might be supposed. In 1936 Gallup underestimated the Democratic Party vote by 6.9 percentage points, in 1940 by 2.5 points, in 1944 by 1.5 points, and in 1948 by 5.0 points. Similarly, Crossley underestimated the Democratic presidential vote by 6.9 per cent in 1936, by 1.6 in 1944 and by 4.7 in 1948. On the other hand, Roper, who had never missed a presidential election by more than one percentage point and who had been within 0.2 of the Roosevelt vote in 1940 and 1944, had the largest error of all in 1948 with an underestimate of the Truman vote of 12.4 percentage points (see Table 1).

Table 1
National Popular Vote * and Predictions

	Actual	Predictions			Error in Percentage Points		
		Gallup	Crossley	Roper	Gallup	Crossley	Roper
Truman	49.5	44.5	44.8	37.1	−5.0	−4.7	−12.4
Dewey	45.1	49.5	49.9	52.2	4.4	4.8	7.1
Wallace	2.4	4.0	3.3	4.3	1.6	0.8	1.9
Thurmond	2.4	2.0	1.6	5.2	−0.4	−0.8	2.8
Others	0.6	0.0	0.4	1.2	−0.6	−0.2	0.6
	100.0%	100.0%	100.0%	100.0%			

* Based upon the final figures compiled by the Associated Press.

Both the polling predictions and the general picture in the public press were misleading, not only in their forecast of a Dewey victory, but in their analysis of the voting trends. The percentage of the votes cast for the Republicans was remarkably close to the percentage achieved in 1944. Governor Dewey polled 45.8 per cent of the national vote in

---

* The Editor solicited this paper and gives it priority in publication because of its importance and timeliness. Only rarely would a situation arise which would justify such special treatment.

1944 and 45.1 per cent in 1948. The reason why the 1948 election was close was not that there had been a gain in the Republican vote, but that there were defections from the Democratic vote to Governor Thurmond and Henry Wallace. Though the national percentage total for Dewey remained constant, there were interesting shifts in the sectional support he received in the two elections. The Republican candidate made slight gains in the industrial east and on the Pacific Coast, but suffered real losses in the west-central states, namely in Iowa, Kansas, Minnesota, Missouri, Nebraska, South Dakota and Wisconsin. Neither the polls nor the newspapers detected this very significant reversal of national voting behavior in which Truman carried a number of the farm states.

State-by-State Errors of the Polls

In their predictions of the specific states the polls almost doubled their average state error of 1940 and 1944. They were not as far off as in 1936, save that their error this time was one of sign as well as magnitude, i.e. they missed the winning candidate. Crossley's average state error of 4.4 was almost a percentage point better than Gallup's. The Crossley poll missed 11 of the 48 states by six percentage points or more as against 16 similar misses by the American Institute of Public Opinion. It is significant that most of Gallup's large errors were in states where the Republicans lost votes from 1944 to 1948. Where the Republicans made gains Gallup's errors tended to be smaller. In other words, the Gallup prediction was that of a general increase, fairly evenly distributed over the nation, rather than a differential increase in certain states. This means that no simple correction for Gallup's inflation of the Republican vote on a state-to-state basis would have remedied his inaccuracies. Table 2 presents the state-by-state errors of the Gallup and Crossley polls. Roper made no state estimates.

General Reasons for the Failure of the Polls

To the world of applied research the poor predictive performance of the polls was as much of an upset as the election of President Truman was to the newspaper world. Yet from a scientific point of view there was evidence, before November 1948, that the polls could not continue their successful record without a change in basic methodological approach as well as in specific techniques. The general philosophy of the pollsters was one of rule-of-thumb procedure rather than sound theory and method. What had worked in the past was accepted at face value without an analysis of why it had worked nor an analysis of the conditions under which it had worked. Moreover, their specific techniques of sampling, of interviewing, of research design were known to have serious weaknesses.

The pollsters began in 1936 with an improvement upon the *Literary*

<div align="center">

Table 2

State-by-State Errors of Gallup and Crossley

</div>

	% of Major Party Vote for Truman *	Error in Percentage Points	
		Gallup	Crossley
Alabama **			
Arizona	53.8	− 0.8	+ 1.2
Arkansas	61.7	− 8.7	+ 1.3
California	47.6	− 4.6	− 3.6
Colorado	51.9	− 2.9	− 2.9
Connecticut	48.4	− 4.4	− 8.4
Delaware	48.8	− 1.8	− 0.8
Florida	48.8	− 3.8	− 0.8
Georgia	60.8	− 2.8	+ 1.2
Idaho	50.0	− 3.0	− 5.0
Illinois	50.1	− 4.1	− 7.1
Indiana	48.4	− 4.4	− 4.4
Iowa	50.3	− 7.3	−10.3
Kansas	44.6	− 5.6	− 2.6
Kentucky	56.7	− 7.7	− 3.7
Louisiana	32.8	+ 6.2	− 4.8
Maine	42.3	− 0.3	− 3.3
Maryland	48.0	− 4.0	− 2.0
Massachusetts	54.7	− 9.7	− 7.7
Michigan	47.6	− 3.6	− 0.6
Minnesota	57.2	−11.2	− 9.2
Mississippi	9.8	+ 5.2	+ 8.2
Missouri	58.1	− 6.1	− 3.1
Montana	53.1	− 3.1	− 4.1
Nebraska	45.8	− 7.8	− 3.8
Nevada	50.4	− 3.4	− 2.4
New Hampshire	46.7	− 2.7	− 5.7
New Jersey	45.9	− 3.9	− 4.0
New Mexico	56.4	− 5.4	− 4.4
New York	45.0	− 6.0	− 3.0
North Carolina	58.0	− 7.0	− 1.0
North Dakota	43.4	− 5.4	− 4.4
Ohio	49.5	− 7.5	− 4.5
Oklahoma	62.7	− 7.7	− 4.7
Oregon	46.4	− 4.4	− 4.4
Pennsylvania	46.9	− 2.9	− 4.9
Rhode Island	57.8	− 3.8	− 4.8
South Carolina	24.1	+13.9	+ 4.9
South Dakota	47.0	− 6.0	−10.0
Tennessee	49.1	+ 1.9	− 1.1
Texas	65.4	+ 0.6	+ 0.6
Utah	54.0	− 4.0	− 6.0
Vermont	36.9	· 1.9	− 5.9
Virginia	47.9	− 3.9	− 1.9
Washington	52.6	− 5.6	− 6.6
West Virginia	57.3	−11.3	− 7.3
Wisconsin	50.7	− 9.7	− 7.7
Wyoming	51.6	− 4.6	− 5.6
Average State Error		5.2	4.4

Digest biased method of sampling. Since 1936 they made some minor improvements and learned either to take advantage of their compensating errors or to correct for their biases, but they never made major advances in methodology. Why, then, did they do so well in 1940 and 1944 with their methods and techniques and so poorly in 1948? The two main reasons seem to be: (1) Their experience with techniques and corrections in Roosevelt elections. With a change in the political scene their procedures no longer functioned effectively. Thus Gallup and Crossley started in 1936 with an error of 6.9 percentage points, improved their performance in subsequent Roosevelt elections, but moved back toward their original starting point when they attempted a presidential election in which different factors were operative; (2) The Roosevelt elections were highly structured situations in which the dominant personality of Roosevelt crystallized attitudes and opinions. With this definite bipolarity of attitude it was not difficult, even with poor techniques, to make election predictions.

Moreover, the polls have never adequately examined the nature of the problem of prediction. In basic science, predictions are made not for an open system of events but in terms of contingent conditions. In applied science, the engineer or the weather forecaster makes some estimates of the possible determinants of the process or event he is attempting to predict. Similarly in attempting to make predictions about social behavior, the social scientist must take into account the relevant field of forces. He cannot merely single out a behavioral or attitudinal trend and predict its repetition. Yet this is essentially what the pollsters attempt to do. They reproduce the national election in miniature and assume that the final election will be a repetition of the trend they have measured without recourse to the many determinants of voting behavior.

It should be emphasized at the start that their fundamental mistake is not to be found so much in any one technique, such as quota sampling or fixed-alternative questions, as in poor research design. In basic science and in applied science we attempt to measure the relationship between two variables and seek to establish causal connections. We do this, moreover, at some level of generality beyond the specific content of one particular situation so that we can build up generalizations which apply to the same type of social process. This means that we do more than report the given percentage of people who favor the Marshall Plan or say that they will vote for President Truman. This means, moreover, that we must conceptualize and identify the important variables and obtain systematic measures of them.

If we apply this logic of research design to election prediction, we need to set up a number of studies designed to measure the determinants

of voting behavior or turn-out and the causal conditions affecting political conviction. It is not enough to have some rough measure of background variables such as income level, or amount of schooling, or even union membership. We need some picture, in addition, of the intervening variables which will give us the perceptions and attitudes related to political parties and political party candidates. How much of this can be done by public opinion polls is a debatable question, but it is scarcely in their best interests to continue to lag behind the advances made by psychologists and social scientists in their studies of human behavior. These points have all been made before the 1948 polling debacle and can be found in the writings of A. Campbell, D. Cartwright, R. Crutchfield, D. Krech and the present writer.[1]

Sources of Error in the 1948 Polls

It will never be possible to make a precise assessment of the contribution of every factor to the error of the polls. Since the polls did not set up adequate hypotheses about voting behavior and political preferences during the campaign, the data are not now available for analysis. It is not even possible to go back and reinterview the same respondents sampled by the polls because the polls did not take names or addresses. There are some limited panel studies where this is being done and they will throw some light upon the problem. Gallup did ask a sample of respondents to return postcards after the election to indicate how they voted, but the selective bias in a mail-return makes these data hazardous to interpret.

It is usually assumed that the important sources of error, however, are to be found in: (1) differential turn-out; (2) the undecided voter; (3) the changing voter; and (4) the representativeness of the sample.

Differential Turn-out

Australia is the pollsters' Utopia, for in Australia the law requires all citizens to vote. It must be remembered that in our country forecasting an election involves two predictions: an estimate of how voters feel about the candidates and an estimate of which voters will go to the polls on election day. In general the polling organizations make no systematic correction for turn-out but depend upon their educational bias in sampling for the major adjustment.

[1] A. Campbell. Polling, open interviewing, and the problem of interpretation. *J. Soc. Issues*, 1946, 2, 67–71; D. Cartwright. Review of G. Gallup's *A guide to public opinion polls*. *J. consult. Psychol.*, 1945, 9, 201–202; R. Crutchfield and D. Krech, *Theory and problems of social psychology*. New York: McGraw-Hill, 1945; D. Katz. Survey technique and polling procedure as methods in social science. *J. Soc. Issues*, 1946, 2, 33–44; and D. Katz. The interpretation of survey findings. *J. Soc. Issues*. 1946, 2, 62–66.

One explanation of both Dewey's defeat and the pollsters' failure is that the Republicans stayed away from the polls in greater numbers than they usually do, as compared to the customary voting behavior of the Democrats. The reasons marshalled to support this theory are varied and not too consistent. For example, the opinion polls defeated themselves by making the Republicans overconfident and so less energetic about getting out the vote; or the Republicans were apathetic about their standard bearer; or the farmers were too busy getting in the harvest on election day to go to the polls.

The hypothesis of Republican overconfidence, or indifference, in its effect upon turn-out, makes sense only if we assume that the polls were accurate in their original estimates about the wishes of the people. It can be argued more plausibly that the nature of the turn-out in 1948 *reduced* rather than *increased* the prediction error. Neither party did a good job on turn-out in 1948. Many Democrats as well as Republicans stayed away from the polls. Against the overconfidence of the Republicans was the lack of motivation on the part of millions of Democrats who idolized Roosevelt and found Truman a weak substitute. Since there are considerably more people in the country who consider themselves Democrats than consider themselves Republicans and since young people who come of voting age are more likely to favor the Democratic than the Republican ticket, the chances are that if the national turn-out had been as heavy in 1948 as in 1940, there would have been a Truman landslide and not a Dewey victory. The overconfidence hypothesis ignores the fact that party machines are organized on a local and state basis. Even though the Republicans thought the presidential election was in the bag, there were many Congressional, state and local offices in doubt, for which it was necessary to turn out the vote. And the states in which overconfidence should have been the highest according to this theory were the states where Dewey actually made gains as in Maine and Vermont.

There is no proof that the upper-income Republican groups relative to lower-income groups failed to vote in greater numbers in 1948 than in the past. The figures in Table 3 show turn-out by economic groups for 1948 and the heavy-voting year of 1940.

The NORC survey in 1940 showed that the lowest income group stayed away from the polls in a ratio of three to one compared to the highest income group. The Roper figures show an even higher ratio in 1948 in favor of greater turn-out among the upper-income people. It should be stated, however, that the comparability of these figures leaves much to be desired. They were obtained by two different organizations and the income groupings may vary considerably. They are suggestive, however, in their implication that the 1948 turnout actually favored

the Republicans.   The same inference was made before the election when experts asserted that a turn-out of under 49,000,000 would help the Republicans.

Table 3

Turn-out by Economic Groups in 1948 and in 1940

Economic Group	Did Not Vote	
	1948 Post-election Poll by Roper	1940 Post-election Survey by NORC
A	11.3%	16.0%
B	14.6 ⎫	32.0
C	26.7 ⎬	
D	40.6 ⎭	47.0

Similar interpretations come from a study of the farm and city vote. If the Truman victory were a matter of differential turn-out, then we would expect bigger Democratic majorities than usual in the industrial centers where the unions and Democratic machines are entrenched. But this was not the case.   Truman lost a number of industrial eastern states and ran surprisingly well in the farm belt and in rural districts. Preliminary analysis of rural and urban counties corroborates the national trend.   Dewey lost not because the Republican farmers stayed away from the polls but because many of them voted for Truman.

Though turn-out does not seem to be the explanation for the difficulties of the pollsters, it is essential in future research that attempts be made to measure, or take into account more thoroughly, the factors which affect turn-out.   Certainly much more can be done to get at the spontaneous forces within voters which get them to the polls on election day.   Crossley has made a start on this problem with questions on voting intention and certainty of voting but in addition we need to study the potency of the individual's involvement in both the national and local elections, the importance the individual attaches to his own vote, and his feeling of responsibility toward voting participation in the democratic process.   The external factors are more difficult to get at but unless we know something about the relative strength of political machines in various states and the pressures of the individual's own social group, we are handicapped in making predictions.

The Undecided Voter

A larger proportion of people than usual could not, or would not, tell interviewers how they were going to vote on election day.   The Roper survey in August, 1948 found 15.4 per cent of the people undecided and Gallup and Crossley still had about 8 per cent undecided in October.

The polling predictions were computed with the *undecided group omitted* on the assumption that these people, to the extent they voted, would distribute themselves among the presidential candidates in the same proportions as the decided voters.

The mistake in this assumption was that the great majority of the undecided were not at a mid-point between the two major candidates. Many people were undecided between Truman and the minor party candidates or between Truman and not voting at all. There is direct and indirect evidence that the undecided vote went more heavily to Truman than to Dewey.

The Survey Research Center of the University of Michigan asked people about their voting intentions in an October study which was being conducted for another purpose than election prediction. The question was asked to get a measure of political identification for correlational analysis. Since names and addresses were available the same panel was re-interviewed after the election and queried about their actual voting behavior. The people who originally said they did not know how they would vote, now reported that they voted for Truman in a ratio of two to one. Fewer of this undecided group reported that they voted than of the decided group. Though the national sample was small, the results are consistent with other findings. Similar evidence will be available from other panel studies.

The indirect evidence comes from an examination of the undecided group in pre-election surveys. Roper's results show that many more of the people, who did not know how they would vote, considered themselves Democrats than considered themselves Republicans. Roper also asked about such issues as rent control, social security measures, and the Taft-Hartley act. The undecided group were not consistent in their responses, but on the whole, they resembled Truman supporters more than Dewey supporters.

The undecided voter was thus one source of the polling error in prediction. But because the undecided group was after all a minority and because they did not turn out to vote as much as the decided group, they could not have contributed more than about one per cent of the five per cent prediction-error.

The failure of the polls to study the undecided vote illustrates the lack of research design in their methods. It would have been possible to have set up systematic hypotheses about this group and explored the nature of their indecision, the reasons for their indecision, their basic political philosophy, etc. The Roper poll had some data on the undecided group but it made no real use of the information it had. In the past the undecided vote was not a problem in election prediction and even in 1948 it may not have been a major factor. Nonetheless it

may loom larger in future elections. But more important is the consideration that it is related in its psychological dimensions to the problem of the changing voter.

The Changing Voter

Another source of polling error was the fact that some people told the interviewer one thing and then behaved differently on election day. This distortion is twofold in nature. Part of the problem is a matter of interviewing skill and technique, in that people may give what seems like a socially acceptable answer in the interviewing situation. If Dewey is supposed to be the popular candidate, if his is the name they ordinarily hear in everyday conversation as the assured winner, and if they have some doubt about Truman, people may find it easier to say "Dewey" when asked the direct question about voting preference. There is no documentation for this possible source of error in the last election but it suggests the importance of thorough interviewing and real training of interviewers.

The second part of the problem concerns genuine psychological change. In a difficult choice-situation some people may give one response to an interviewer but when confronted with the reality of the election booth they may change their minds.[2] Take, for example, the supporter of the New Deal, dissatisfied with Truman, who says before election he will vote for any candidate save Truman. When the chips are down, however, he returns to the party most representative of his beliefs. Another type of change is typified by the farmer who originally planned to vote for Dewey, became alarmed at the fall in farm prices and the Republican position on the support of farm prices, and voted in terms of what seemed to him his best self-interest.

Panel studies give some support to the changing voter as a source of polling error. More people who said they would vote Republican, after the election report they voted Democratic than report the reverse. It is not possible to estimate precisely how much this was responsible for the prediction failure. In a post-election survey (see Table 4), Gallup found that in general Dewey voters had made up their minds earlier in the campaign than Truman voters.

These findings indicate that the 1948 political situation had a different psychological structure for many people than the Roosevelt elections. In the final analysis most people may have voted for the party which represented their welfare as they saw it. But they did not crystallize their beliefs until they had to. They finally reached a decision consonant with their basic attitudes. This may be why so many people who talked against Truman were so delighted with the election returns.

[2] This hypothesis about the changing voter was suggested by R. Crutchfield.

Table 4

Gallup Post-election Survey of Time Voters Made Up Their Minds

Definitely Made Up Their Minds	Truman	Dewey	All voters
Before campaign started	46%	64%	54%
Early in campaign	11	12	12
First half, Oct.	4	2	3
Second half, Oct.	13	5	9
Election day	5	3	4
Indefinite	21	14	18
	100%	100%	100%

The lesson for election prediction, presented by the undecided and changing voter, is not primarily the necessity of polling until the last moment. Trend studies must be made, but adequate research in this field should be more than the projection of a single attitudinal trend. The polls can interview 48 hours before the election and still miss the voter who reacts differently to the reality of the election booth than to the straw ballot. The real lesson is that the determinants of political behavior must be systematically explored. We need to study how the voter perceives political parties and candidates; for example, to what extent is he politically-minded in viewing a candidate and a party as an instrument for protecting and improving his interests, to what extent is he reacting to the personalities of the candidates, etc. We need, further-more, to investigate the basic social, economic, and political beliefs and their relative importance to him.

To do thorough studies of this kind requires much more theoretical planning than the polls have thus far done. They occasionally ask questions on issues, but they have not systematically designed studies to give answers to problems of political motivation. These studies, whether conducted by the pollsters or by psychologists, are indispensable to the making of predictions. In addition to better research planning, the use of intensive interviewing, even on a pre-test basis, could get the signifi-cant frames of reference in which people are thinking. The usual polling pre-test is one of testing question-wording, not one of the experimental investigation of the dimensions of the problem under study. Lazarsfeld has pointed out how adequate pre-testing with intensive interviewing could be used to develop more valid ballots with pre-coded answers.[3]

The Representativeness of the Sample

To estimate turn-out, to allocate the undecided vote, to gauge the stability of voting preference all require good interviewing and research

[1] P. Lazarsfeld. *The controversy over detailed interviews. Publ. Opin. Quart.*, 1944, 8, 38–60.

design which goes beyond the direct question of voting intention into the related causal factors. In addition, however, there is the problem of sampling, of obtaining a truly representative cross-section of the electorate. The quota-control method of the polls has been under fire for some time and since it is a more palpable weakness than lack of study design, the controversy over polling methods will focus unduly about it.

The quota-control method sets up a cross-section which in theory represents the larger population proportionately in terms of sex, age, socio-economic status, urbanization, and geographical area. Interviewers are assigned quotas on this basis and told to bring back results from respondents of given characteristics. It is sometimes contended that the quota-control method is vulnerable because it does not stratify on some variable related to voting behavior such as union membership. This argument fails to get at the essential weakness of quota-control sampling. If the cross-section obtained by the quota method really achieved a random representation of the population according to the controls it employs, the chances are all in favor of other characteristics such as religion, occupation and even union membership being properly represented.

The real defect of the quota-control method is in its execution. Since there are no strict controls over interviewers, they in fact select the sample. The result is not a random, or true probability sample. Interviewers are told to bring back results from so many respondents in the D, or *below-average economic category*. What constitutes a D respondent and how D respondents are to be selected is too much a matter of interviewer judgment. In practice interviewers filling their quotas take people who are physically and psychologically more accessible. As middle-class members themselves they under-represent the poorer people and to some extent the very wealthy. Since the wealthy are much less numerous, these are not compensating errors. Moreover, interviewers tend to get respondents more like themselves on other counts than would be found in a truly representative sample.

The under-representation of the lower income groups is in evidence whenever a quota sample is broken against some measure indicative of socio-economic status such as education or telephone ownership. Uncorrected quota samples employing no special devices to limit interviewers traditionally find between 12 to 20 per cent too few people in the lower education brackets.

In 1940 and 1944 Gallup and Crossley corrected indirectly for the quota bias by adjusting for past voting behavior. In 1948 Gallup also used the respondents' answers to questions on education to correct his final sample. In spite of corrections Gallup and Crossley never succeeded in the Roosevelt elections in eliminating their Republican over-*estimates*.

Roper does not employ corrections but stands by the raw data from
his sample. In the Roosevelt elections he had the advantage of sizable
compensating errors in that his southern sample was much too Demo-
cratic and his northern sample much too Republican. In 1940, for ex-
ample, Roper overestimated the Democratic vote in the East South
Central states by 12.5 per cent and in the South Atlantic states by 8.3
per cent. To balance this, however, he underestimated the Democratic
strength in the West North Central states by 6.8 per cent and in the
Mountain states by 10.5 per cent. In 1948 the candidacy of Governor
Thurmond knocked this compensating error into a cocked hat. In some
southern sections Roper interviewers found Truman and Thurmond
tied. Without his usual southern overweighting of the Democratic vote,
there·was nothing to compensate for the northern Republican inflation
and Roper after three highly accurate predictions was not even close in
1948. If his figures are corrected for the educational bias in his quota
sample, however, the Roper figures are very much like the Gallup and
Crossley predictions.

It is clear, then, that the largest part of the Roper error was due to
the uncorrected quota-control method of sampling. It is not clear,
however, how much of the remaining error in prediction (about five per
cent) is due to poor sampling, which cannot be corrected for, and how
much to non-sampling factors. Those who defend quota sampling admit
that area, or probability, sampling would have given slightly greater
accuracy but they dismiss sampling as a minor factor. Their logic on
this score is interesting in that their final argument is that quota sampling
costs less than area sampling.

That poor sampling did contribute to the polling error in spite of the
corrective adjustments made in the data seems a sound interpretation
for these reasons:

1. Even in the Roosevelt elections, with a highly structured situa-
tion, with attitudes and opinions well crystallized before election day,
Gallup and Crossley were not able to correct away their inflation of the
Republican totals. In 1940 Gallup missed 16 states by three percentage
points or more, but only one of these errors was in the direction of over-
estimating the Democratic vote. In 1944 he was off the mark by three
percentage points or more in 22 states. Only two of these errors were
overpredictions of the Democratic vote. Similarly, Crossley had 13 state
errors, in 1944, of three percentage points or greater, but only one of these
favored the Democratic candidate.

2. Corrective adjustments introduced into data to compensate for
poor sampling are always limited by the poor sampling that was done in
the first place. To inflate an under-represented group by some corrective

weight does improve the whole sample to some extent relative to the neglected group, but it cannot insure the representativeness of this group. If, for example, we weight up the people with no better than grade school education by fifteen per cent, we still have not improved the character of the sample for this group even though we have improved its place in the sample.

3. The area method of sampling was more accurate in the 1948 elections than the quota sample but it was used in too limited a way to draw definite conclusions. The study of the University of Michigan Survey Research Center, previously referred to, used a national area sample and found an even division among the decided voters between Truman and Dewey. This evidence is limited in that the sample was small and in that the Center's methods of interviewing also differ from polling methods. The Elmira study of Lazarsfeld, using an area sample, missed the vote in that town, however, by six per cent. Gallup's quota sample for New York state was also six per cent in error. The clearest evidence is from the University of Washington Survey group which tried both area and quota sampling for the state of Washington. The area sample had an error of 2 percentage points; the quota sample an error of 7 percentage points.

The interpretation of this evidence is obscured by the fact that neither method of sampling was followed according to its literal requirements. In the case of the quota method, the interviewers who used it were new to this method of sampling. They reported that they were unable to fill the lower income quota in 20 per cent of the cases. Whether this is the usual difficulty with the quota method which happened to be reported here because of the newness of the interviewers or whether this is unusually inadequate quota sampling is a matter of debate. The area sample also was not carried out perfectly and utilized liberal substitutions. Nevertheless, the final figures show a superiority of the area sample over the quota method of five percentage points.

Though it is unlikely that all of the prediction error was due to quota sampling, it may have contributed between one and three percentage points to the Gallup and Crossley underestimations of the Truman vote. If this estimate is correct, then area-sampling would have indicated a closer election and counselled caution on all-out predictions.

Applied Psychology and the Polls

The growing criticism of the polls in the field of applied psychology has already become more sharply focussed with the 1948 prediction failure.[4] Criticism has been directed at two main phases of polling operations: (1) the failure of the polls to keep abreast of technical and methodo-

logical advances in pure and applied social psychology or to do methodological research of their own; and (2) the reluctance of the polling agencies to make public their data and their procedures such as sample size, the exact corrective adjustments employed, etc.   Both of these points were made by the technical committee of social scientists, serving the Congressional Committee which investigated Gallup in 1944.

This criticism is undoubtedly justified but it should not lead to a blanket condemnation of the public opinion polls.   They have made real contributions in the past in stimulating a quantitative and factual approach to problems once dealt with by journalistic reporting or armchair political science.   They can make greater contributions in the future if they take stock of their methods.   The 1948 setback should be of real value to them in that they may see that they have been hampered by a blind empiricism in the past.   This empiricism led them to feel that what had apparently worked once or twice or even three times was somehow sacred and could be relied upon to work in the future no matter how conditions changed.

Nor should the failure of the public opinion polls be construed as an indictment of all research in the field of consumer needs and wants. There are many studies in this field to which the weaknesses of polling techniques do not apply.   As in any new research, standards and methods in measuring consumer reaction vary considerably.   It is of interest that some market research organizations, interested in sound methodological development, had accepted true probability sampling before November 1948.

Though social psychologists in general may want to work for improvements in polling methods, it is important to distinguish between the polls and basic research in social psychology and the social sciences. Field work employing quantitative measurement on social psychological problems should not be confused with polling any more than the laboratory work of the psychologist on problems of perception should be confused with market research.   Though there are areas of overlap the tendency of the layman to confuse basic and applied research should not mislead the professional worker.   This does not mean that the social scientist should be completely divorced from applied research.   It is his responsibility to help formulate standards of research that will help both types of research.   Such standards are needed in the public interest and by the polls themselves.   They cannot afford a repetition of their 1948 experience.

Note on the Cardall Practical Judgment Test

Dorothy H. Carrington

Institute for Psychological Services, Illinois Institute of Technology, Chicago, Illinois

The Cardall Practical Judgment Test (2) was given to over 300 unselected men who had come for vocational guidance to the Illinois Institute of Technology. Their age range was from 16 to 63 years and the educational level from 8th grade to persons holding the Ph.D. degree. All subjects also took: the Adams-Lepley Personal Audit Test (1); the ACE Psychological Examination (1942, college edition) (5); and the Otis Gamma (4). Scores on each of these tests were correlated with the Practical Judgment scores. The scores of the Practical Judgment Test were also correlated with age and education.

The results are given in Table 1.

Table 1.

Correlations between Practical Judgment Test and Other Variables

Variables	No. of Cases	r	Standard Error of r	Significance at: 5%	1%
Age	361	.02	.0528	NS	NS
Education	349	.21	.051	S	S
ACE Total	310	.29	.046	S	S
ACE Quantitative	311	.24	.053	S	S
ACE Linguistic	307	.29	.052	S	S
Otis Gamma	344	.37	.046	S	S
Personal Audit					
Seriousness-Impulsiveness Scale	315	−.05	.0563	NS	NS
Firmness-Indecision Scale	316	.20	.054	S	S
Tranquillity-Irritability Scale	315	−.03	.0564	NS	NS
Frankness-Evasiveness Scale	316	.10	.0559	NS	NS
Stability-Instability Scale	313	.15	.0567	S	S
Tolerance-Intolerance Scale	314	.02	.0565	NS	NS
Steadiness-Emotionality Scale	307	.17	.055	S	S
Persistence-Fluctuation Scale	309	.09	.0567	NS	NS
Contentment-Worry Scale	309	.07	.0568	NS	NS
Split Half r corrected by Spearman Brown Formula	275	.69	.013		

The split half reliability of the scores on the Practical Judgment Test in the sample used is .69 corrected by the Spearman Brown Formula. There is a low but statistically significant positive correlation between the Cardall Test and intelligence as measured. The correlation with formal education is also low but significant. For the most part, the Cardall Test does not correlate with the sub-parts of the personality test used although on Firmness-Indecision, Stability-Instability, and Steadiness-Emotionality there is a slight positive correlation which is significant. For the sample studied, the Cardall Test scores are independent of age.

Cardall correlated his test with the Army Alpha, Link's Personality Quotient, Bell's Adjustment Test and college grades and found no significant correlations. His correlation of the Practical Judgment and intelligence was −.05. This differs by 42 points from the results obtained in the present study. He does not give the number of persons in his sample nor the type of population on which they are based so it is difficult to tell what factors are causing the discrepancy.

The data in this study differ from the data in Cardall's Manual (2) in that in the Illinois Institute of Technology group, practical judgment as measured is not totally independent of intelligence and academic background, and there is an indication that some personality factors influence test scores. According to the sample studied the reliability of the test is too low for the test to be used for individual predictions.

References

1. Adams, C. R. *Manual of directions for the Personal Audit Test.* Chicago: Science Research Associates, 1945.
2. Cardall, A. J. *Manual of directions for the Practical Judgment Test.* Chicago: Science Research Associates, 1942.
3. Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill Company, 1942.
4. Otis, A. S. *Manual of directions for Otis Quick-Scoring Mental Ability Tests.* Yonkers-on-Hudson: World Book Company, 1937.
5. Thurstone, L. L., and Thurstone, Thelma G. *Psychological examinations for college freshmen.* Washington, D. C.: The American Council on Education, 1942.

Originality Ratings of Department Store Display Department Personnel

Catherine P. Dougan, Ethel Schiff and Livingston Welch

Institute for Research in Clinical and Child Psychology, Hunter College

In this study we attempt to measure creative thinking of the employees in the display department of R. H. Macy's, by means of the Welch Reorganization Test (1, 2) which obliges the subject to recombine familiar ideas according to four different patterns. It is Welch's (1) assumption that the ability to recombine easily and reorganize ideas according to a specific plan is essential to all types of creative thinking. His contention is not that this is the only factor involved, but the individual lacking this ability will be seriously handicapped in an imaginative capacity.

In two previous studies the Reorganization Test was given to 30 professional artists, 25 art majors and 48 unselected students. We will compare the results of these investigations with those obtained in the present study.

Procedure

1. *The Reorganization Test.* The test is divided into four parts. The first three sub-tests make use of written material and the fourth makes use of blocks. The total testing time is 26 minutes.

Part 1. Instructions

Recombine the words of each group on the next page to make as many meaningful grammatical sentences as possible. For example, here is a group of ten words,

MEN SKY IS FIGHT THAT THE SLOW BRIGHT OF FOR

which can be recombined in the following sentences:

> Men fight for the sky.
> The sky is bright.
> The fight is slow. Etc.

You will receive as much credit for a *short* sentence as for a long one. Your sentences do not have to be artistic, but they must be grammatical. There must be at least a subject and a predicate. You will receive credit for a sentence which is only slightly different from another. A word from the group can be used only once in the same sentence, but it may be used any number of times in other sentences. Only use words from the group that you are examining at the time. You may skip from one group to another, if you like.

There are ten of these groups and you have only ten minutes in which to complete the test. Are there any questions? . . . Do not turn the page until the examiner says "Start."

The following are the ten groups of Part 1:

1. Dog tree climbs runs those a smooth good by with
2. City John built stood a that large strong of from
3. Car fence travels was this that big cool for by
4. Sea woman move could these the green rough with of
5. Den lion ate is big deep these the of by
6. House child left has blue frightened the a for by
7. Lemon wife cooks finds that soft round with from
8. Potatoes maid cut once small hot these a of for
9. Fish boy waits catches the a long cold by from
10. Slowly the golden light that rested upon them moved away

Part 2. Instructions

Make as many letters as possible using no more and no less than three straight lines. For example, the letter A is made with three straight lines, two slanting downward and one across. You will be given no credit for the letter A, since it is an example.

Make as many letters as possible, using no more and no less than two straight lines.

Make as many letters as possible, using no more than one straight line and one semi-circle.

The time limit is three minutes.

Part 3. Instructions

On the next page you will be given a list of twenty words which you are to connect into a story. You must be certain to use the words in the order in which they appear on the list. If the first word is "tree" in your story this must be the first word which appeared on the list. You must not skip any of the words.

Your story must be grammatical and logically related. It must have a beginning and an end. You will be rated on the number of words you make use of in the time allotted. Write as fast as you can and underline each of the twenty words as you use it.

The time limit is three minutes.

The words used in this test were:

STAIRS OCEAN CHEMISTRY SONG TEST MOUNTAIN BUBBLE DOG LEMON PICTURE POST BLANKET VIOLIN LAMP NIGHTMARE STEAM LEG WINDOW SWAMP STAMP
(The words were given in this order.)

Part 4. Instructions

The object of this test is to construct out of ten blocks on each trial as many pieces of furniture or home furnishings as possible. The pieces of furniture you construct must fit properly. It must be symmetrical and be recognizable as a piece of furniture. Do not attempt to be futuristic. Use conventional forms. You must use a minimum of two blocks to construct a piece of furniture. You can make as many of the same type of furniture as you like. You will receive full credit for the same type that is only slightly different from another.

You have only ten minutes to complete this test. There are five trials. Hence, you have only two minutes for each trial.

The blocks used in all five trials were geometric shapes selected from a box of playing blocks. On each trial the blocks were presented to the subject on a piece of cardboard with each shape outlined so that the positions of the blocks were standardized. A record was kept of all of the combinations of blocks for which credit was given.

2. *Rating Scale.* Each subject in the display department was rated on a five point scale by the manager and by the assistant manager of that department. These men rated their employees independently; however, it must be borne in mind that these two men, over a period of time, must have exchanged ideas as to the originality or creative thinking of their employees.

Subjects

In the present study the Reorganization Test was given to a total of 33 employees in the display department of R. H. Macy's. Their individual positions ranged in talents and included: artists, window, show-case and floor display men, designers, stylists, and the executives in charge.

Results

The test results of these 33 department store employees in the display department were compared with those of the three groups, 30 professional artists, 25 art majors, and 48 unselected students, reported in the previous study. The mean scores and standard deviations for each group on each part of the test are shown in Table 1.

Table 1

The Mean Performance Scores and the Standard Deviations
for Each Group on Each Part of the Test

Parts	Professional Artists N = 30		Art Majors N = 25		Display Personnel N = 33		Unselected Students N = 48	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	17.7	7.2	21.9	7.6	10.8	6.8	18.0	4.2
2	12.5	1.9	13.2	1.0	11.9	2.3	6.7	1.8
3	11.4	4.1	7.3	2.5	6.8	4.0	9.1	3.2
4	18.4	7.8	13.9	9.1	14.0	10.5	3.4	2.7
Total Score	60.5	12.3	56.4	15.1	43.1	18.1	37.6	7.0

It will be seen that the display personnel compare almost equally with the unselected student in total score, whereas both are considerably lower than the art majors and professional artists. The only sub-test in which there is a striking difference of score is in Part 4, which is concerned with the construction of furniture with blocks. It is interesting to note, therefore, that a large part of the department store

personnel tested were employees of the furniture and interior decorating departments.

The difference between professional artists, unselected students, and art majors has already been mentioned in the previous studies.

All of the differences between sub-tests were put to test and some significant t-values were obtained. The t-values obtained for differences between the means of the four groups on each part of the test and between the total score are presented in Table 2.

Table 2
The t-Values Obtained for Differences Between Means of Groups*

Parts	Prof. Artists and Art Majors	Prof. Artists and Unselected Students	Art Majors and Unselected Students	Display Per. and Prof. Artists	Display Pers. and Unselected Students
1	2.1	0.2	3.0	4.0	6.0
2	1.5	13.2	17.1	.4	11.5
3	4.4	1.1	2.4	4.5	2.8
4	2.0	10.2	7.5	1.9	0.8
Total Score	1.1	10.9	7.2	4.5	1.91

* $t_{05} = 2.0$; $t_{02} = 2.3$; $t_{01} = 2.6$.

It appears that, for the total test score, the difference between the display personnel and the professional artists is statistically significant, while that between the display personnel and the unselected students is not. Parts 2 and 4 of the test seem especially important. In all cases except between the display personnel and the unselected students, the differences between the groups on these two parts seem to be consistently significant.

In order to determine the degree to which the test results agreed with the creative ratings given by the department managers, the test scores and the ratings were analyzed. A coefficient of .60 was obtained.

It was considered that perhaps Part 4, alone, would be of high enough reliability for judging creative thinking. However, when the results of Part 4 were analyzed it was found that the P value for this sub-test when used singly was small enough to cast doubt on the hypothesis.

Summary and Conclusions

The purpose of this study was to measure originality of department store display personnel. A special test was constructed by Welch in which the subjects were obliged to recombine familiar ideas according to a

series of four different patterns. The subjects were rated by their supervisors by means of a 5 point rating scale to provide performance data with which to correlate the Reorganization Test results.

1. A contingency coefficient of .60 was obtained between the ratings given by the manager and the scores resulting from the Reorganization Test.

2. The performance of the display personnel was compared with that of subjects examined in a previous study. The mean scores for the four groups are as follows: professional artists, 60.5; college art majors, 56.4; department store display personnel, 43.1; and unselected students, 37.6. The difference between the professional artist and the display personnel is statistically significant while that between the unselected student and the display personnel is not. However, the display personnel were superior to unselected students on two of the sub-tests.

3. These results indicate that there is a possibility of measuring originality, as it would apply in the fields of advertising and display.

Received July 12, 1948.

References

1. Welch, L. Recombination of ideas in creative thinking. *J. appl. Psychol.*, 1946, 30, 638–643.
2. Welch, L., and Fisichelli, V. R. The ability of college art majors to recombine ideas in creative thinking. *J. appl. Psychol.*, 1947, 31, 278–282.

The Rosenzweig Picture-Frustration Study in the Selection of Department Store Section Managers

H. Wallace Sinaiko

L. Bamberger & Co., Newark, N. J.

This report is an outgrowth of a study of certain intelligence and personality characteristics of Department Store Section Managers.[1] A battery consisting of two tests of mental ability and one measure of personality was administered to a group of 53 of 58 employed Section Managers.

Findings with regard to the two intelligence tests were essentially negative: correlations between test scores and a quantitative rating of job performance were so low as to be chance deviations from zero. Similar treatment of scores from the personality test—the Rosenzweig Picture-Frustration Study—produced more fruitful results in terms of predicting job performance in Section Managing.

Method

The Instrument. The P-F Study consists of 24 cartoon-like pictures in booklet form. Each picture illustrates a frustrating situation involving two or more people. One figure is shown saying something about the situation while the caption box over the second person is blank. The subject is told to write the first reply that comes to his mind in the blank over the person being addressed in the picture.

Six principal scores are derived from the P-F Study. Responses are categorized according to *direction* of aggression and *type* of reaction. "Direction" categories include the following: (1) Extrapunitiveness,—aggressions directed by the subject toward someone or something in the frustrating situation; (2) Intropunitiveness,—aggressions directed by the

---

[1] A definition of the Job "Section Manager" appears in the Dictionary of Occupational Titles, Part I, page 576, as follows: "*Manager, Floor;* aisle man; floorman; manager, section; (retail trade); 0-75.10; supervises employees in a designated section of the selling floor; instructs new workers and sees that they follow store system in making sales; shifts selling personnel from one department to another so that service will be efficient and prompt; regulates lunch hours and grants permission for employees to leave the floor; handles returned goods, approves bank checks, and adjusts claims or refers them to the adjustment department; answers customers' questions relative to merchandise or location of merchandise; floor-walker (almost obsolete)."

subject toward himself; (3) Impunitiveness,—the absence of aggressive feeling. "Types" of reaction are: (1) Obstacle-Dominance,—the problem, or situation, is predominant in the subject's response; (2) Ego-Defensive,—blame, or responsibility, is assigned for what has happened; (3) Need-Persistence,—a *solution* to the problem is mentioned. The scoring categories are symbolized by the letters E, I, and M, each corresponding to the "direction" of aggression. "Type" of reaction is signified by the use of the symbols O-D, E-D, and N-P.

The Group. As mentioned above, 91% of the 58 Employed Section Managers comprised the experimental group. Breakdown of the group by sex was: 44 women and 9 men, 83% and 17% respectively. Length of time on the job ranged from three months to 16.7 years (median = 18 months, $Q_1 = 6.5$ months, and $Q_3 = 66$ months). Formal education ranged from eight years to seventeen years. One-eighth of the group had not completed high school, 50% had completed one or more years of college, and 15% were college graduates. Age ranged from 21 to 57 years (Median = 30, $Q_1 = 24$, and $Q_3 = 36$).

The Criterion. A quantitative measure of job performance was built with information obtained from Executive Personnel History forms. This is a modified linear rating scale used throughout the company in its semi-annual personnel review and rating of all executives. Executives, or Section Managers in this case, are rated on six basic qualities, each being subdivided into from two to ten categories. These qualities include Character, Intelligence, Intuition, Experience, Adaptability, and Special Skills. Ratings are made on the subdivisions under each main category. For example, Intuition is rated for each of two points: "Are decisions based on limited data usually correct?" and "Are decisions arrived at without undue delay?"

Actual ratings assigned to the subdivisions are confined to the following values: Outstanding, Above Average, Average Plus, Average, Below Average, and Unsatisfactory.

Each Section Manager is rated by one of four Floor Superintendents. All ratings are checked by the Chairman of Personnel Reviews. Thus, there is a "common denominator" roughly operating to keep ratings by different supervisors comparable. All Section Manager ratings used were on a minimum of three months' service on the job.

To convert the above descriptive ratings into quantitative terms arbitrary weights were assigned as follows: Outstanding, 11; Above Average, 9; Average, Plus 7; Average, 5; Below Average, 3; and Unsatisfactory, 1. Mean point values were computed for each of the six basic qualities weighted and summated.

A seventh basic category on the Executive Personnel History form,

"Placement and Development," was treated slightly differently. The subdivisions, "Is he well placed on his present job?", and "Is he satisfied with his present status?", were weighted as follows: Yes, 10; Yes, qualified 5; No, −10; and No, qualified −5. This weighted score was algebraically added to the summated averages of the preceding six rated qualities. This final figure gave us a quantitative criterion measure for each Section Manager. A frequency distribution of ratings for the entire group showed a range of approximately 50 points (24.6 to 74), a mean of 58.9, and a standard deviation of 9.6.

Results

Ratings of each Section Manager's job performance and the six principal scoring categories of the P-F Study were correlated (Pearsonian product-moment r). Table 1 summarizes these relationships as well as those between P-F Study scores and length of service.

Table 1

Pearson Correlations between Rosenzweig Picture-Frustration Study Scores, Length of Service, and Job Ratings of 53 Department Store Section Managers

	Picture-Frustration Study Scores					
	E	I	M	O-D	E-D	N-P
Length of Service	−.23	.15	.39**	.11	−.12	−.01
Job Ratings	−.31*	.28*	.25	−.02	−.48**	.38**

* Significant at the 5% level.
** Significant at the 1% level.

Discussion

Length of Service. The distribution of this variable had marked skewness toward the right, or longer period of time on the job. The mean length of time in Section Managing was approximately 34 months while the median was only 18 months. Hence, there seems to be a fairly high rate of turnover, with only a small group of "long stays" in the job.

One score on the P-F Study showed a statistically significant relationship to length of service: Impunitiveness. Thus, there is a tendency among the longer-staying Section Managers to show more M (minimizing, absence of blame-placing, conformity) in their responses than in the more recently hired of the group.[2]

[2] Correlations were run between age and P-F Study scores. One significant relationship, $r = -.28$, ±.05, was found between Extrapunitiveness and this variable. All other correlations between age and P-F Scores approximated zero.

Job Ratings. There were four statistically significant correlations between P-F Study scores and job ratings. Keeping in mind the requirements of a Section Manager's duties, these relationships follow a logical pattern. There was a negative correlation between the criterion and E: $r = -.31$, $\pm .05$. Better Section Managers show relatively fewer extrapunitive, aggressive, responses. Management requires of its Section Managers a constant display of good-will in their customer contacts. A large number of these contacts occur under strained circumstances produced by such things as complaints about quality of merchandise, non-delivery, or service, etc. Section Managers must obviously refrain from any show of aggressiveness in handling these adjustments if they are to maintain customer friendship toward the store.

The correlation between the criterion and I, $r = .28$, $\pm .05$, indicates that better Section Managers show a tendency to turn their aggressions against themselves. Intropunitiveness may be a necessary adjunct of efficient Section Managing. The somewhat hackneyed phrase, "The customer is always right," is an attitude actually encouraged by Management. In other words, store policy regarding customer relations is itself an intropunitive one.

M scores were correlated with the criterion to a positive, but statistically insignificant, degree: $r = .25$, $\pm .10$. There is a slight tendency for better Section Managers to avoid defining situations as conflictual and to see them as non-frustrating.

The first of the P-F Study scores relating to *type* of reaction, O-D, showed a practically zero correlation with the criterion: $r = .02$, $\pm .50$.

The highest correlation between P-F Study scores and job ratings was found between E-D and the criterion: $r = -.48$, $\pm .01$. Thus, low-rated Section Managers tended to be more defensive when confronted by the test situations; i.e. they were overly concerned with fixing responsibility, either in assuming blame themselves or in blaming someone else.

N-P scores on the P-F Study showed a moderate relationship with the criterion: $r = .38$, $\pm .01$. High-rated Section Managers tend to have an adaptive, or solution-seeking, attitude for dealing with every-day problem situations.

Additional Statistical Data. Correlations were run between ratings and four variables, length of service, age, education, and sex. This was done to determine whether any of the reported relationships might be an artifact of one of these variables. Correlations were as follows: (1) between age and ratings: $r = .19$, $\pm .15$; (2) between length of service and ratings: $r = .34 \pm .02$; (3) between education and ratings: $r = .25 \pm .06$; (4) between sex and ratings: $r = .27 \pm .04$. Thus, the

H. Wallace Sinaiko

latter three variables, length of service, education, and sex, are related to job ratings to a statistically significant degree. Women tended to get higher ratings than men.

P-F Study scores of the 15 highest-rated Section Managers are compared with scores of a like number of the lowest-rated in Table 2. The comparison of mean P-F Study scores of top-rated and bottom-rated Section Managers, shown in Table 2, confirms the earlier discussed correlational findings. However, the statistical significance of these differences is greatly reduced by the small number of cases. In any event, differences *do* exist in the direction indicated by the overall correlations on the total group of 53 Section Managers.

Table 2

Comparison of P-F Study Mean Scores* of the 15 Highest-Rated Section Managers and the 15 Lowest-Rated Section Managers

	Scoring Categories					
	E	I	M	O-D	E-D	N-P
Highest-Rated						
Mean	37.1	32.5	30.6	18.1	48.5	33.6
Sigma	17.05	7.5	11.8	7.2	8.2	10.9
Lowest-Rated						
Mean	43.4	29.6	27.0	17.4	54.0	28.1
Sigma	14.7	7.6	9.6	5.9	9.7	12.0
t	1.07	1.15	.90	.59	1.83	1.28
p	.28	.24	.36	.50	.06	.20

* Values for each category represent the *proportion* of the total number of responses made in the test falling in that category.

Table 3 compares quartiles of P-F Study scores of the 15 highest-rated and 15 lowest-rated Section Managers. That there is a great deal of overlap between the top-rated and bottom-rated Section Managers' scores is apparent. Thus, the P-F Study is not a highly valid selection device by any means, although tendencies do seem to be indicated insofar as performance in Section Managing is concerned.

A further check on the efficiency of the P-F Study with the present occupational group was made by using a "combined P-F index."[3] The index was built by adding the *number* of I, M, and N-P responses made by each Section Manager, and then subtracting the number of E and E-D responses. In this way a simple algebraic expression, which could

[3] This index was suggested to the writer by Dr. H. G. Gough, Department of Psychology, University of Minnesota, in a personal communication.

Table 3

Comparison of P-F Study Quartiles* for 15 High-rated
and 15 Low-rated Section Managers

	Scoring Categories																	
	E			I			M			O-D			E-D			N-P		
	Q_1	Md	Q_3	Q_1	Md	Q_3	Q_1	Md	Q_3	Q_1	Md	Q_3	Q_1	Md	Q_3	Q_1	Md	Q_3
High-rated Group	21	34	48	23	34	44	17	32	40	12	19	23	41	50	54	22	33	44
Low-rated Group	32	43	52	24	29	35	19	26	35	10	18	23	52	54	66	23	26	35

* Values for each category represent the *proportion* of the total number of responses made in the test falling in that category.

be either positive or negative, was derived for each of the top-rated 15 Section Managers and each of the bottom-rated 15. A comparison of indexes thus obtained on each of the two groups of Section Managers is shown in Table 4.

If a cutting score were to be established at plus 2 we would eliminate 5 of the top-rated 15 Section Managers and 11 of the bottom-rated 15.

Table 4

Comparison of Combined P-F Study Indexes of 15 Highest-Rated and 15 Lowest-Rated Section Managers

Case	Indexes	
	High-Rated	Low-Rated
1	16.0	15.0
2	15.5	12.0
3	12.5	7.5
4	12.0	6.0
5	12.0	0.0
6	9.0	−1.5
7	6.0	−2.0
8	4.5	−4.5
9	4.0	−6.0
10	2.5	−6.5
11	−4.0	−6.5
12	−7.5	−12.5
13	−10.0	−13.0
14	−13.5	−17.0
15	−18.5	−27.0

The use of a simple index, such as that described here, corroborates the discussion of Table 3. Thus, the P-F Study is far from being a highly valid selection tool although it does warrant some consideration in the hiring of Department Store Section Managers.

Summary

1. The Rosenzweig Picture-Frustration Study was administered to 53 Department Store Section Managers. Quantitative measures of job efficiency were built from personnel review data and correlated with each of the six principal scores derived from the P-F Study.

2. Statistically significant negative relationships occurred between the criterion and scores for Extrapunitiveness, and between the criterion and Ego-Defensive scores. Positive, statistically significant relationships between the criterion and Intropunitiveness, and between the criterion and Need-Persistive scores were found. A positive, but not significant, correlation was found between the criterion and Impunitiveness. A near-zero relationship existed between job ratings and Obstacle-Dominance scores on the P-F Study.

3. A simple technique of combining P-F scores into an index would admit 10 out of 15 top-rated Section Managers and would reject 11 out of 15 bottom-rated Section Managers if a cutting score of plus 2 was used.

4. This investigation suggests that the Rosenzweig Picture-Frustration Study measures factors which are associated with occupational success as a Section Manager, and which might have value in an employment selection program.

The Rorschach as a Predictor of Academic Success

Boyd Rowden McCandless

Ohio State University

Many studies have been made, and many claims advanced for the Rorschach as a highly useful test in the area of academic prediction. The thinking behind the studies is perhaps best summarized in Klopfer and Kelly (3, p. 266):

"If the Rorschach method could do nothing else but estimate the intellectual level of the subject as well as the usual intelligence tests, these tests would be preferable since they are simpler to apply. The importance of the Rorschach method for the intellectual aspect of personality diagnosis lies in something which no intelligence test attempts, the differentiation between potential capacity and actual efficiency."

Beck (1), Rappaport by implication (5) and Munroe (4) in a careful experimental study of college women concur in such an estimate of the Rorschach test as a measure of intelligence and a predictor of success.

Munroe (4) has worked out a 28 item check list, usable with either the group or individually administered Rorschach. It is filled in by a *protocol inspection method*, and general adjustment has been found by her, working with women students at Sarah Lawrence, to correlate negatively with number of checks accumulated by the subject. In general, girls with fewer than 10 checks were reasonably adequately adjusted; girls with more than 10, moderately to seriously maladjusted (4, p. 66). The "Inspection Rorschach" adjustment rating predicted academic success somewhat better than did ACE percentile ratings, coefficients of contingency .43 and .36 for 348 subjects; corrected, .49 and .39 for the two tests respectively, (4, p. 76).

Beck (1) has devised an organization, or Z score, to be derived, essentially, from individually administered Rorschach tests. To quote:

. . . the sum of all the Z scores in any Rorschach record is the measure of S's organization activity. These totals vary directly as the intelligence of S. The Z factor has certain virtues not inherent in W. For one thing, it takes account of much activity that W misses. Second, since it is not scored in discrete units, as is necessary in the case of W, it makes it possible to take account of intermediate values and continuous distributions, and is thus a more flexible measure. Third, it is an index of the intellectual energy as such, irrespective of the *kind* of intelligence that S uses, something that does influence W. Thus Z is a more accurate representative of the intelligence functioning per se. . . . it is therefore an index of thinking power. Its essence is the capacity to grasp relations not perceived by others (1, p. 12).

The three authors, Beck (1), Klopfer and Kelly (3) and Rappaport (5), less directly than the first two, assign predictive values in a general fashion to many categories of the Rorschach. Munroe (4), as stated, has done so specifically and empirically.

Of the score for number of whole responses to blots produced, Beck says: "The higher the intelligence potential of an individual, the more W he can produce" (1, p. 10). Klopfer and Kelly state: ". . . (W) represents an emphasis on the abstract forms of thinking and the higher forms of mental activity" (3, p. 259). Both qualify the quantitative use of this W score, stating that the quality of W must be considered.

Of large detail (D) and small detailed responses to the individual cards, Beck states that where emphasis is on D there is revealed "a person who attends to obvious and practical interests" where Dd shows an "evidence of some need to pursue too much the elements that most people disregard"; and emphasis on W "is the sign of an over-all thinker" (1, p. 14). Klopfer and Kelly believe that the individual with approximately ⅔ of his responses listed as D and Dd "has enough common sense to use the most obvious material before he starts seeking the unusual" (3, p. 260).

Summarizing the thinking of the various authors on other categories, with perhaps some injustice, it appears to be, from the point of view of making predictions of efficiency:

Animal (A) responses tend to indicate a certain amount of conformity of thought, too few indicating unusual thought processes, too many barrenness, lack of creativity and stereotypy.

Popular responses have roughly the same meaning.

Percentage of responses made on the basis of form alone indicate in a general way an intellectual, unemotional approach to life; the percentage of form responses at a superior level is directly related to functioning intelligence.

Human movement responses betoken creative imaginativeness, with qualifications set on their location and type.

Responses dominated by color, but with form present (CF), betoken an adjustment intermediate between infantile and fully, socially adult, as far as emotional control is concerned. Responses dominated by form but using color are given by the emotionally fairly rich but controlled, mature person.

Vista or perspective responses (FV or FK) are used by persons who are self-critical and liable to "inferiority feelings." Flat grays (FY or FC') used in responses to the cards indicate anxiety and reduction of intellectual energy.

The broader the subject's interests and the richer his educational back-ground and the higher his intelligence, the greater the variety of things he will see.

In general, the more intelligent and the less anxious the subject is, the more complete human figures he will see; and the more whole human and animal with relation to detailed human and animal figures there will be in his record. Finally, the more intelligent he is, the larger the total number of responses he will give. Klopfer and Kelly (3, p. 208), however, do not agree fully with this. The seeing of things in the white spaces tend to betoken resistive, persistive and unusual methods of approach.

It is with these elements of the Rorschach that the author has concerned himself in this study. He realizes most clearly that the Rorschach is essentially a configurative test, where a pattern of factors must be taken account of to make any really adequate interpretation or prediction. On the other hand, he feels that the more checks made on the predictive efficiency of the specific categories of the Rorschach for which predictive efficiency has been claimed, the more valid is the use of the test for such purposes of prediction. In the case of this study, this prediction is in the areas of academic progress and achievement.

Subjects and Method

Individual Rorschach's were given in conjunction with vocational guidance, during the writer's assignment as Selection and Classification Officer to the U. S. Maritime Service Officers School, Alameda, California to approximately two hundred Officer Candidates. These men were aspiring for marine licenses and commissions in the U. S. Maritime Service, and undergoing a four months' period of training pursuant to that end. Every subject who could be matched on the eight criteria used was selected for this study.[1]

These men were "normal" in that they were functioning adequately in a wartime society, contributing to the war effort, making in general adequately and highly motivated progress toward their specific goal, and were in no case undergoing psychiatric treatment.

Thirteen pairs of men were matched on the basis of AGCT score; average Mechanical Comprehension Test score; average Iowa Silent Reading comprehension test score (form Am, new, advanced); average Stanford Advanced Arithmetic Reasoning Test score; average age and amount of education; marital status (six married, two divorced, five single in each group); and enrollment in division of the school (ten

[1] This study reflects the author's conclusions and is not an official Maritime Service publication.

members of each group were enrolled in Deck training, three in Engine training).

The basis of differentiation was in terms of the academic grade averages. With a value of 5.0 assigned to grade A and 1.0 assigned to grade F, the high grade point group averaged 4.7 ranging from 4.5 to 5.0; and the low grade point group struggled through the school with average grades of 2.9, ranging from 1.0 to 3.6. Some, indeed, of the low grade point group failed to qualify academically for their licenses and commissions.

Table 1

Quantitative Characteristics and Differences between
High and Low Grade Point Groups

Characteristic	High Grade Point	Low Grade Point	Diff.	t of Diff.	Significant at % level
AGCT	135.7	135.7	0.0	0.000	Greater than 5
MCT	141.5	139.4	2.1	0.430	Greater than 5
Arithmetic Equated Score	94.1	89.5	4.6	1.367	Greater than 5
Reading Standard Score	103.9	99.4	4.5	1.213	Greater than 5
Age (years)	25.6	24.4	1.2	0.679	Greater than 5
Education (years)	12.3	12.1	0.2	0.605	Greater than 5

It will be noted from Table 1 that these men are very superior, psychometrically speaking. The mean AGCT score for the rank and file was set at 100, with a S. D. of 20 points. The average for this group was 1.75 S. D. above the national mean.

The groups average 2 S. D. above the mean on mechanical comprehension, as measured; and in math and reading comprehension, approach the average third year college man.

The ranges for the equating scores are given in the following tabulation:

Variable	High Grade Point	Low Grade Point
AGCT	124–149	124–150
MCT	123–161	119–161
Math	67–103	65–105
Reading	91–113	93–120
Age	19–37	19–32
Education	10–16	10–14

Between these two groups of men, so similar in quantitative characteristics, so different in academic success, the problem was to distinguish, if possible, personality characteristics which might explain the efficiency differences.

Their Rorschach's were studied intensively in an effort to find such distinguishing characteristics.

Results

The results of the present study were negative, with one exception. Table 2 summarizes the averages for the high and the low grade point groups. Differences in the direction of the low grade point group are indicated by a minus (−) sign; t (Edwards (2)) is given in the fourth column and the level of significance of the t in the fifth column.

It will be noted that t approaches the one per cent level of confidence in only one case,—mean number of popular responses. Even here, the difference is of little practical significance (8.1 versus 6.6 mean popular responses for the respective groups).

Favoring the high grade point group with t's above 1.0, are found for: Mean number large detail responses; Mean number tiny detail responses; Mean number space responses; Mean number human movement responses; Mean number pure form responses; Mean number superior pure form responses; Mean number animal responses; Mean number human detail plus animal detail responses; Mean number popular responses; and Mean quality of whole responses.

Favoring the low grade point group with t's above 1.0, are found for: Mean number whole responses; Mean number achromatic color responses (both including and excluding texture); and Mean number of color-form responses.

It has been considered by most of the authors working with the Rorschach that the ratio of W (whole blot) responses to the number of M (human movement) responses is one of the best of the predictive factors for "efficiency" or productivity, with a ratio of 3 to 1 being considered optimal. Ratios falling materially below 3.0 are considered to characterize "underproductive" persons; ratios falling materially above 3.0 are considered to characterize "over-striving" persons, whose performance is likely to be describable as "quantity" rather than "quality." These latter may produce much, be over-ambitious, under considerable strain; and their products are likely to be superficially acceptable rather than really good.

If such considerations hold for these two groups, we should expect the high grade point men to have a mean ratio approximating 3.0, which, if deviant, would probably be expected to be above 3.0; the low grade point group would be predicted to show a mean ratio falling below 3.0. As can be seen from Table 2, the opposite is true, the high grade point men showing a mean ratio of 1.6; the low grade point men a ratio of 3.4. The difference, however, is not a statistically significant one.

Beck's Z or organization score (a measure of the "capacity to grasp relations not perceived by others" (1, p. 12)) differentiated even less effectively than the conventional Rorschach categories discussed above.

Boyd Rowden McCandless

Table 2
Selected Rorschach Differences and their Significance
for High and Low Grade Point Groups

Mn. for Category	High Grade Point	Low Grade Point	Diff.	t for Diff.	% Level of Confidence
N Responses	39.4	32.6	6.8	0.916	Greater than 5
N Whole R's[1]	6.4	10.0	−3.6	1.532	Greater than 5
N Detail R's[1]	26.1	19.8	6.3	1.284	Greater than 5
N Tiny Det. R's[1]	6.1	2.7	3.4	1.183	Greater than 5
N Main and Additional Space R's[1]	11.3	8.4	2.9	1.029	Greater than 5
N Human Mov't R's	5.8	3.5	2.3	1.257	Greater than 5
N Animal Mov't R's[2]	4.2	3.2	1.0	0.957	Greater than 5
N Inanimate Nov't R's[2]	1.8	1.7	0.1	0.121	Greater than 5
N Vista R's	1.9	1.8	0.1	0.146	Greater than 5
N Form R's	18.8	13.1	5.7	1.528	Greater than 5
N Superior Pure Form R's[1]	13.8	10.3	3.5	1.701	Greater than 5
N Superior of Total R's in form[1]	30.4	25.4	5.0	0.935	Greater than 5
N Achromatic Color R's[2]	2.2	4.8	−2.6	1.525	Greater than 5
N Achromatic Color R's[1]	3.2	4.9	−1.7	1.068	Greater than 5
N Form-texture R's[2]	3.0	2.3	0.7	0.321	Greater than 5
N Form-color R's[4]	3.2	3.4	−0.2	0.119	Greater than 5
N Color-form R's	1.5	2.5	−1.0	1.308	Greater than 5
Mn. Sum Color[2]	3.4	4.5	−2.1	0.764	Greater than 5
N Human R's	3.5	2.6	0.9	0.956	Greater than 5
N Animal R's	15.5	12.1	3.4	1.555	Greater than 5
N Human Detail + Animal Detail R's	10.8	4.0	6.8	1.789	Greater than 5
N Anatomy R's	2.5	2.0	0.4	0.610	Greater than 5
N Popular R's[1]	8.1	6.6	1.5	2.836	Less than 5 (1)
N Response Categories	11.3	10.6	0.7	0.741	Greater than 5
Z Score[1]	43.3	48.5	−5.2	0.409	Greater than 5
N Checks Munroe	11.4	12.1	0.7	0.350	Greater than 5
N Superior Wholes[3]	1.8	2.4	−0.6	0.555	Greater than 5
Mn. Whole Quality[1]	2.0	1.7	0.3	1.863	Greater than 5
N Whole::N Human Mov't Ratio[5]	1.6	3.4	−1.8	0.691	Greater than 5
Human + Animal R's:: Human Detail + Animal Detail Ratio	2.7	5.8	−3.1	0.591	Greater than 5
Human Mov't::Sum Color Ratio[5]	2.0	1.1	0.9	0.408	Greater than 5

[1] After Beck's (1) criteria.

[2] After Klopfer's and Kelly's (3) criteria.

[3] After Rappaport's (5) criteria.

[4] There were too few pure color, or texture-form or pure texture responses to compute a legitimate difference.

[5] Based on 11 pairs, due to zero in numerator of 2 ratios. t was computed on these ratios as with N's, since the various Rorschach authors seem to regard the relationship as a unit or entity.

What slight, statistically non-significant discrimination it did make was in the wrong direction; mean Z score for the high grade point men was 43.3, with a range from 9 to 99.5; for the low grade point men, mean Z score was 48.5 range 6 to 122.5. t was .409 for this difference.

As a final check, Munroe's (4) check sheet, which gave positive results for the Sarah Lawrence students, was filled out for each man. Here the small difference was shown in the right direction (high grade point men averaged 11.4 checks; low grade point men 12.1 checks). The range was wider, however, for the former group (4-28) than for the latter (4-20). The men would appear to be seriously maladjusted, also, according to Munroe's findings, who considers ten checks as a cutting score (4, p. 66). Her students were not given the individual Rorschach, which may account for the greater number of checks earned by this group.

Discussion

Despite the consistently negative results of this investigation, certain trends appear according to prediction. In general, the high grade point men are seen to be slightly more controlled emotionally or with less emotion to control, slightly more productive; on most criteria, slightly less anxious. They tend to show up with higher averages, even when the factor of their higher productivity is cancelled out, in the scores which indicate conformity (except for space responses), and appear, although not significantly, better able to attend to the large, usual; and the tiny, unusual details of the Rorschach blots. If one can generalize from such a tendency, it might be said that such a solid, conforming, non-theoretical approach is one of the bases for academic success, particularly in a "cram" type of program such as the Officer Candidate programs tended to be. The only significant difference (more popular responses for the successful students) fits this trend.

The author does not feel that the findings of this paper detract from the clinical use of the test; but he believes it essential that many such checks as this be made. Finally, he grants the extreme difficulty of the task to which the Rorschach has been set in this case (restricted range and high level of ability, possible similarity of personality due to choice of occupation, small number of cases, etc.). Many authors, however, appear to have taken it for granted that the task could easily be accomplished.

Finally, other patterns, or combinations of the factors discussed above, or some total scoring, weighting system other than Munroe's could conceivably be found to make a clear differentiation between these groups of men who differed so significantly in performance in the highly moti-

vated Officer Candidate situation. The author's repeated scrutiny of the tests has failed, however, to reveal such patterns.

Summary

Two matched groups of Officer Candidates, U. S. Maritime Service, who differed widely in academic achievement in a highly motivated, wartime, officer training program, were given individual Rorschach's with the following results:

. 1. An analysis of the conventional Rorschach categories failed to demonstrate any important statistically significant differences, although trends appeared.

2. Munroe's (4) check list which discriminated good from poor students at Sarah Lawrence college failed to show differences in this group.

3. Beck's (1) Z or organization score also failed to make discriminations. In fact the latter showed slight mean differences in a direction opposite to expectations. The statistically non-significant, but consistent trends were toward more emotional control, more conformingness, less anxiety on most criteria, more attention to concrete details, and slightly greater productivity for the high grade point men.

References

1. Beck, S. J. *Rorschach's Test, II.* New York: Grune and Stratton, 1945, Pp. xii + 402.

2. Edwards, A. L. *Statistical analysis.* New York: Rinehart and Company Inc., 1946, Pp. xviii + 300.

3. Klopfer, B., and Kelly, D. McG. *The Rorschach technique.* New York: World Book Company, 1942. Pp. x + 436.

4. Munroe, R. L. *Predictions of the adjustment and academic performance of college students by a modification of the Rorschach method.* Stanford University: Stanford University Press, 1945. No. 7 of the Applied Psychological Monograph. Pp. 104.

5. Rappaport, D., Gill, M., and Schafer, R. *Diagnostic psychological testing.* Chicago: Year Book Publishers, Inc., 1946. Pp. xi + 516 (Vol. II).

The OL Key of the Strong Vocational Interest Blank for Men and Scholastic Success at College Freshmen Level *

Stanley R. Ostrom

Department of Public Instruction, Dover, Delaware

Psychologists have developed instruments that measure abilities and aptitudes with a fair degree of accuracy. The use of these instruments for prediction purposes in learning situations has not proved as successful as one might hope, however. This may be due, to some degree, to non-intellectual traits which cause some individuals to persevere through discouragements while others of apparently equal potential fail. The measurement of these traits has proved most elusive.

Counselors using the Strong Vocational Interest Blank for Men have to a large degree assumed that the Occupational Level key of the Strong blank is one approach to this problem. This position is verbalized by Darley (2, pp. 66):

> Clinical experience together with limited experimental data would indicate that the lowest occupational level scores on the revised blank will accompany the interest type previously defined as "lower level jobs." Furthermore, an excessively low occupational level score seems at present to be associated with lack of "staying power" or "survival power" in college competition. This hypothesis should be tested as quickly as research data accumulate, by careful studies of matched groups, since it is a phase of the "level of aspiration" and general motivational problem.

Strong holds the same position stating "Men with high OL scores have the interests of business executives and professional men, but those with low scores have the interests of workmen" (5, pp. 195). He further suggests that the key has value for a counselor helping a student plan his high school or college training program (5, pp. 203-204).

Specific statistical studies for the corroboration of these hypotheses are, however, very meager. Berdie (1) reports a correlation of only .03 between the OL key and academic achievement of forty-three college students. He also found an equally low correlation, .01, when he compared the OL scores with curricular satisfaction.

Stanley R. Ostrom

Kendall (3), on the other hand, obtained positive results when he studied 300 male college freshmen in Syracuse University. He found that when academic ability as measured by the Ohio State Psychological Examination, Form 21, was held constant three groups distinguished by differing levels of OL were found to differ in college achievement. His three groups consisted of 100 men each of high, average, and low OL. The difference between these groups when adjusted for ability by covariance proved significant beyond the five per cent level but not to the one per cent level of confidence. Kendall concluded "if used with caution OL scores at the extremes of the distirbution should be helpful to the counselor in making judgments concerning individual chances for scholastic success."

These studies give impetus to the need for further research as suggested in the last sentence of the statement by Darley referred to above.

To test further the above hypothesis the writer conducted a study in which an attempt was made to determine the relationship between the OL key of the Strong Blank and scholastic achievement at three levels of education. The following discussion is a report of the findings at the college freshman level.

As is the case each year, the 1946–1947 freshman class at Syracuse University participated in a testing program shortly after enrolling in school. Among other tests taken by the men were the Ohio State Psychological Test, Form 21 and the Strong Vocational Interest Blank for Men. From these test data six groups of seventy-five men each were chosen according to the following criteria:

High level, high ability: Men whose OL scores were equal to a standard score of fifty-seven or above and whose raw scores on the Ohio State Psychological Examination, Form 21 were ninety and above.

Average level, high ability: Men whose OL scores were between standard scores of forty-seven and fifty-two, and whose raw scores on the Ohio State Psychological Examination, Form 21 were ninety and above.

Low level, high ability: Men whose OL scores were equal to a standard score of forty-five and below, and whose raw scores on the Ohio State Psychological Examination, Form 21 were ninety and above.

High level, low ability: Men whose OL scores were equal to a standard score of fifty-seven or above, and whose raw scores on the Ohio State Psychological Examination, Form 21 were below ninety.

Average level, low ability: Men whose OL scores were equal to a standard score of between forty-seven and fifty-two, and whose raw scores on the Ohio State Psychological Examination, Form 21 were below ninety.

Low level, low ability: Men whose OL scores were equal to a standard score of forty-five and below, and whose raw scores on the Ohio State Psychological Examination, Form 21 were below ninety.

Findings

The mean honor point ratios were determined for each of the six groups. From Table 1, it can be seen that an even step progression from

low to high OL and from low to high ability emerged except in one instance, that of average to high OL in the low academic group.

Table 1

Average Honor Point Ratios for Six Groups of Syracuse
University Male Freshmen (Total = 450)

	Mean Honor Point Ratios		
	High OL	Average OL	Low OL
High Ohio	1,742	1,569	1,357
Low Ohio	1,058	1,104	1,036

These data were then subjected to analysis of variance. Table 2 shows F-ratios for both OL and academic aptitude at magnitudes great enough to justify the rejection of the Null Hypothesis at the one per cent level of confidence.

Table 2

Analysis of Variance: Multiple Classification for 450
Syracuse University Male Freshmen
(Determining Effects of Ability and Level)

Source of Variance	Degree of Freedom	Sum of Squares	Mean Squares	F *	Test of Hypothesis**
Ability	1	238,496	238,496	72.23	Reject *
Level	2	37,971	18,985	5.75	Reject
Interaction	2	28,323	14,162	4.27	. . .
Residual	444	1,467,232	3,302	. . .	. . .
Total	449	1,772,022	. . .	. . .	. . .

* Where F = greater mean square/lesser mean square. By referring to Snedecor's tables of F (4, pp. 222–225), we may use the following three rules in testing the hypothesis: (a) reject the hypothesis tested, if the calculated value of F is greater than the 1% point given in the tables; (b) accept the hypothesis tested, if the calculated value of F is less than the 5% point given in the tables; (c) remain in doubt, if the calculated value of F lies between the 5% and 1% points given in the tables.

** The Hypothesis tested is a null hypothesis concerning the difference between means of groups, i.e., there is no significant difference between the means of groups. (The 1% point necessary for rejection of the Null Hypothesis was 6.70 for ability and 4.66 for level.)

Conclusions and Recommendations

1. A very significant relationship was established between honor point ratio and both academic aptitude and OL in the Syracuse University

freshmen sample. This result strengthens Kendall's study and gives a strong case to the use of OL scores in prediction of college success. It does not, of course, justify the use of the key as a single measure of motivation, but it does point up its rightful place in a predictive battery.

2. Standardization of OL on a school population. The Occupational Level Key of the Strong Blank was standardized by contrasting "unskilled men" and "business and professional men earning $2,500 and upwards a year" (5, pp. 185). An obvious result of using such a scale on a college population is the large number of high OL scores among college students. Finding men from the freshman class for the two low OL groups was extremely difficult. So difficult, in fact, that it was necessary to include men with scaled scores of forty-five to assure groups of seventy-five. Setting up an OL key standardized on college groups would undoubtedly result in a sharper instrument.

3. Follow-up study of college freshmen group. Repeating the college freshmen study four years after the original study will be revealing if the four year college honor point ratios are available for each group.

4. Study of the high OL-low ability college freshmen. No reason is available to explain the sharp drop in mean honor point ratio between the average OL and high OL groups of low ability. An intensive study of a generous portion of this group to find answers for this deviation from the expected pattern is recommended.

References

1. Berdie, R. F. Prediction of college satisfaction and achievement. *J. appl. Psychol.*, 1944, 28, 239–245.
2. Darley, J. G. *Clinical aspects and interpretation of the Strong Vocational Interest Blank.* New York: The Psychological Corporation. 1941.
3. Kendall, W. E. The occupational level scale of the Strong Vocational Interest Blank for men. *J. appl. Psychol.*, 1947, 31, 283–287.
4. Snedecor, G. W. *Statistical methods.* Ames, Iowa: Collegiate Press Inc., 1946.
5. Strong, E. K. *Vocational interests of men and women.* Stanford, California: Stanford University Press, 1943.

Note On the Shifts of Interest with Age

E. L. Thorndike

Professor Emeritus, Columbia University

Thirty-seven men, all graduate students of education, ranging in age from 23 to over 40, reported, as well as they could estimate, the relative strength of the following tendencies, each for himself at the present time, and for himself at the age of 12: Approval (having people look up to you); Mastery (being boss); Kindliness (seeing people happy); Gregariousness[1] (being with one's own crowd); Studying things; Studying people; and Studying abstractions.

These men had been studying educational psychology and had a certain common basis for their definitions of the above. Doubtless, however, the terms did not mean quite the same things to the different individuals, and it would probably be impossible to define with precision just what they did mean to the average of the group. Within limits, however, these terms do have a community of meaning to them and to the readers of this note. The change from 12 to adult age (around 30 in the case of this group) was: a loss of $2\frac{1}{2}$ steps for Approval; a loss of $1\frac{1}{2}$ steps for Mastery; a gain of $\frac{1}{2}$ step for Kindliness; a loss of 2 steps for Gregariousness; a loss of $1\frac{1}{2}$ steps for Studying things; a gain of $3\frac{1}{2}$ steps for Studying people; and a gain of $2\frac{1}{2}$ steps for Studying abstractions. For a group of lawyers, or doctors, or engineers, or business men, the shifts with age might well be different.

These facts seem worth noting, especially the different effect of age upon the interest in studying things as compared with studying people and abstractions, and the absence of any substantial change in kindliness. According to traditional fiction, a boy of twelve is brutal and careless of others.

These same records can be studied from the point of view of the permanence of the tendencies as reported. Assuming the validity of the testimony, the facts show that a person's nature at 12 is prophetic of his nature in adult years in this respect (the median correlation for the 37 cases is $+.55$). The child to whom approval is more cherished than mastery is likely to become a man who seeks applause rather than power, and similarly throughout. The effect of chance errors, forgetfulness, and the like, is to make this correlation too low. The effect of a constant error whereby a person projected his opinion of himself to form his opinion of his own past would be to make the relation closer than it really was. The net result of eliminating these errors would, I conjecture, be to raise the correlations somewhat.

[1] This perhaps would be more suitably named "a mixture of gregariousness and sociability."

A Fallacy in the Use of Median Scale
Values in Employee Check Lists

Clifford E. Jurgensen

Minneapolis Gas Light Company

Several investigators (1, 2, 4) have published articles using the Thurstone equal-appearing intervals method, or a slightly modified form of the method, to select and weight items in a check list to be used for rating employees. The author has developed similar unpublished check lists and is familiar with a number of other unpublished scales developed for or by various companies. It thus appears that the procedure is sufficiently used to warrant mentioning a fallacy which appears when the equal-appearing interval method is used in an industrial merit rating scale.

Briefly, the method consists of obtaining a large number of statements which relate to good or poor job performance. Statements are printed separately on cards which are then sorted by a large number of judges according to the method of equal-appearing intervals. In some cases statements are printed serially and judged by encircling a number from 1 to 9 preceding each statement, this procedure having been shown (5) to give the same results as sorting. The median and semi-interquartile range for each statement is computed by formula or by nomograph (3). Statements with a large semi-interquartile range are eliminated, and the remaining items form a pool from which scale items are selected in such manner that statements differ in scale value by approximately equal differences. A tentatively selected scale is used experimentally, tests of item relevancy are made, and the scale is modified where necessary. The final scale is used by asking raters to check items which describe or apply to the employee being rated. The "score" is the median or mean scale value of the checked statements.

The scaled statement technique assumes that all items form a single continuum which is factorially pure. This assumption has not even been loosely approximated in any employee merit check list seen by the author. The typical employee check list contains items dealing with work output, quality, learning ability, job skills, personality, work habits, and many other types of items. Customary tests of item relevancy are generally applied to statements in such check lists, but these tests eliminate only those items which have a low or negative correlation for the *group* of persons under consideration. It is quite possible that items may show a high positive correlation within a group, but an individual may nevertheless differ widely from the group tendency. For example, studies

show a relatively high positive correlation between speed and accuracy of work. It is not uncommon, however, for an industrial supervisor or executive to challenge this finding on the basis that some of his workers show such great differences in speed and accuracy that the overall finding is untenable to him.

Table 1

Comparison of Two Types of Scale Values
with Reference to Three Employees.

Item	Median Scale Value	Revised Scale Value	Employee		
			A	B	C
Is one of the best employees in the department	8.6	3.6	X	X	X
Has unusually good quality	8.4	3.4	X	X	X
Carries through on all jobs	8.2	3.2	X	X	X
Is extremely loyal	8.0	3.0	X	X	X
Gives close attention to instructions of supervisor	7.8	2.8	X	X	X
Plans work well	7.6	2.6	X	X	X
Has Good judgment	7.4	2.4	X	X	X
Learns new work easily	7.2	2.2		X	X
Is enthusiastic	7.0	2.0		X	X
Reacts favorably to corrections	6.8	1.8		X	X
Starts work earlier than others	6.6	1.6		X	X
Is a steady worker	6.4	1.4			X
Gets help when in difficulty	6.2	1.2			X
Profits from past mistakes	6.0	1.0			X
Is pleasant and courteous	5.8	.8			X
Does fair share of work	5.6	.6			X
Does not alibi when corrected	5.4	.4			X
Total Score based on: Median Scale Value			8.0	7.6	7.0
Revised Scale Value			21.0	28.6	34.0

For purpose of illustration, Table 1 gives seventeen items in order of their scale value as determined by one hundred supervisors. The items form the positive or favorable half of a scaled check list. They are all satisfactory for use in a check list so far as tests of relevancy and ambiguity are concerned.

Ratings of three hypothetical employees are given in columns headed A, B, and C. (It is assumed that none of these three persons has been checked on any items falling below the median scale value of 5.0.) The median scale values for the three employees are 8.0, 7.6, and 7.0 respectively. It will be noted that A is the "best" employee because he does not learn new work easily, is not enthusiastic, does not react favorably to corrections, etcetera! Employee C is the worst of the three employees because he possesses all the listed virtues and performs all the favorable actions!

From a theoretical position it can be contended that the above findings will not commonly be found in actual cases if items are properly selected. However, the presence of the error of median scale score was originally found by the author when it was noticed that the "better" of two employees obtained the lower of the two scores on a scale developed by the usual approved techniques. Other such cases were subsequently found. The decreased validity of the scale (whether large or small) is only one of the objections to the method. An even more serious objection is that the entire scale might fall into disrepute and discard if a few of the raters were to discover that overall scores would increase in magnitude if some of the favorable (but low value) items were not checked even if applicable to the employee being rated.

A simple solution to the above fallacy is to replace median values by positive and negative values obtained by subtracting five from each of the item medians. (This assumes that scaling was based on nine equal-appearing intervals. The constant would differ for other numbers of groups.) The merit rating "score" for each employee is the algebraic sum of the revised weights for items checked as applying to the employee. For the three hypothetical employees referred to in Table 1, the revised scale scores would be 21.0, 30.0, and 34.0. It will be noted that the order of merit is the reverse of that obtained from the median scale value method, and that the revised order is consistent with logic.

Previous discussion has been limited to median scale values. Exactly the same situation, however, is true for mean scale values.

The above is proposed as a simple solution to the error of median scores. The scoring of scaled check lists on the basis of algebraically summed deviations is just as easy as use of mean scale values. Even though the validity of the scale may not be increased greatly (for the group as a whole) by this change in scoring procedure, scores of specific individuals sometimes change appreciably. The use of the inaccurate median (or mean) scale value does not appear defensible on logical grounds even though it seldom results in significant error.

References

1. Ferguson, L. W. The development of a method of appraisal for assistant managers. *J. appl. Psychol.*, 1947, 31, 306–311.

2. Knauft, Edwin B. Construction and use of weighted check-list rating scales for two industrial situations. *J. appl. Psychol.*, 1948, 32, 63–70.

3. Jurgensen, Clifford E. A nomograph for rapid determination of medians. *Psychometrika*, 1943, 8, 265–269.

4. Richardson, M. W., and Kuder, G. F. Making a rating scale that measures. *Person. J.*, 1933, 12, 36–40.

5. Seashore, R. H., and Hevner, K. A time-saving device for the construction of attitude scales. *J. soc. Psychol.*, 1933, 4, 366–372.

An Empirical Approach to a Problem of Psychophysical Scaling *

William H. Angoff

*Human Engineering Branch, Special Devices Center,
Port Washington, New York*

Since Thurstone's original work in the scaling of crimes by means of a paired-comparison procedure (2), numerous psychological judgments, including those concerning attitudes, have been ordered to continua in similar fashion with success. Specifically in industry, a scale has been developed for the quality of work performed by industrial supervisors (3). Another application in industry of the paired-comparison technique may conceivably be that of scaling jobs within an industrial plant. In this latter instance, job levels may be determined by the combination of scale values for factors attaching to a particular job, or they may be determined by the simple scaling of the jobs as a whole without regard to separate factors. In either instance, the jobs could be ordered to a single continuum which would then define the hierarchy. The theoretical defensibility and practical simplicity of such a job evaluation approach appears to constitute unquestionably an advantage over the procedures currently in use. However, there would appear to be a practical difficulty in the situation involved in job evaluation. Wherever job hierarchies are determined, it is frequently the case that new jobs are added or old jobs changed as the plant continues to function. If an entirely new scale of n items or jobs is developed each time the original items or jobs are altered or increased in number, the procedures of scaling, particularly where large numbers of judgments are involved, can become a very costly and time-consuming affair. It is therefore suggested that new items, whether they are jobs or other judgment-objects, may be inserted and placed, as they appear, in their proper positions in a scale that has already been set up and found to be satisfactory; and that a new rescaling of all items would not then be necessary.

The present study attempts to duplicate in miniature such a situation as might obtain in an industrial plant, where a new item is added to a scale which has already been determined, and is presumably in use.

Procedure

A group of ten male movie actors were chosen who are well known to the public, and were used as object-choices. The subjects making the comparisons were twelve in number, and relatively advanced in terms of level of education, intelligence, and sophistication with regard to tastes in moving pictures. The ages of the subjects ranged from 26 to 36 years.

Forty-five cards were prepared with every one of the ten names of the object-choices paired with every other of the remaining names. No pair occurred more than once in the deck of 45 cards. The stimulus-statement was prepared in advance and read by each of the twelve subjects prior to making his choices. The statement read as follows: "In the following pairs of movie actors, choose the one you would prefer to see in a moving picture. Use whatever basis you please for your decision." The choices for the 45 pairs were made separately and independently by each subject and recorded by the experimenter on the spot.

It may be noted that no attempt was made in the experiment to assure uni-dimensionality or high reliability in the eventual scale. The movie actors chosen for the study are all current popular favorites, and it was expected that there would be much disagreement among the subjects with regard to their preference-choices—as indeed there was. Thus as much opportunity as possible was provided to permit the scale values to be affected by the withdrawal or insertion of items. Also, as was expected, the range of scale values that resulted was narrow, permitting slight shifts in scale values to exert considerable effects upon the correlation coefficients that were to be computed.

With regard to the question of uni-dimensionality, it was felt that while the concept is a highly important one in the usual scaling problem, it was a consideration not relevant to the problem here. The purpose of the present study was to manipulate preference-judgments as they were turned in by the judges. The particular manner in which the scale was constructed, and the assumptions underlying the construction of scales were felt to be matters for separate consideration.

The choices having been made by all the subjects, a table of paired frequencies was drawn up, and a standard-score scale-value was determined for each movie actor directly from the percentages. The percentages represented the ratio of his "preferred" frequencies to the total possible "preferred" frequencies. Constant values were added to each scale value to convert them to positive numbers, and finally a 10-item scale was constructed which then constituted the basic or "criterion" scale.

The specific problem now involved deriving a nine-item scale consisting of all but one of the items. This nine-item scale would correspond

to the scale referred to above that "has been determined and is presumably in use." The tenth item, not included in the scale, would correspond to the new item which must be inserted into the pre-existing scale. The question arises: Can we have our judges make paired comparisons only between the new item and the nine old items in order to secure a scale value for this new item; or is it necessary to rescale all ten again? That is, will the information from $n - 1$—in this case, nine—paired-comparisons give as good a scale as the information derived from $n(n - 1)/2$—here, 45—comparisons?

Table 1

Proportionate Frequencies of Preference of Row
Object to Column Object

	Actor										Scale Values
	A	B	C	D	E	F	G	H	I	J	
A	.500	.417	.250	.250	.333	.417	.333	.833	.250	.250	190
B	.583	.500	.500	.500	.583	.417	.500	.750	.417	.583	705
C	.750	.500	.500	.417	.667	.667	.833	.833	.500	.667	962
D	.750	.500	.583	.500	.583	.667	.583	.917	.583	.750	986
E	.667	.417	.333	.417	.500	.583	.417	.750	.500	.667	685
F	.583	.583	.333	.333	.417	.500	.333	.667	.333	.667	559
G	.667	.500	.167	.417	.583	.667	.500	.917	.417	.833	791
H	.667	.250	.167	.083	.250	.333	.083	.500	.167	.167	000
I	.750	.583	.500	.417	.500	.667	.583	.833	.500	.667	875
J	.750	.417	.333	.250	.333	.333	.167	.833	.333	.500	433

By erasing one at a time the columns and corresponding rows in Table 1, ten new scales of nine items each were developed, each time, of course, omitting one of the actors. The scale-values in each of these new scales were then different from the scale-values in the criterion scale, since they had been constructed without consideration of the cells corresponding to the actor omitted in each instance. At this time the scale-value for the omitted actor in each scale was determined independently on the basis of the number of times he was preferred to the other nine. It is apparent that now his percentage value, and consequently his scale-value, was the same as in the criterion scale, since the same number and kind of comparisons were computed for him here as had been computed for the criterion scale. But since his relative scale status was changed because of the changes in the scale-values of the other nine, his relative scale-value was accordingly changed.

The foregoing procedure of drawing out one object-choice at a time and re-inserting into the scale of nine was then modified to answer the

following question: If single-item insertion in a scale of nine results in a ten-item scale that is not substantially different from the scale that would have resulted had all ten items been originally considered at one time (i.e. the criterion scale), then how many items is it possible to insert in a scale before the new scale shows an appreciable departure from the criterion scale? To this end, six of the ten actors were chosen randomly and withdrawn in combination from the scale—first actor X, then actors X and Y together, then X, Y, and Z together, and so on. After each withdrawal, a scale was constructed of the remaining actors, and the withdrawn actors were then reinserted into the scale.

There were two ways in which these insertions could be made. When r actors were withdrawn from the original set of n actors, a scale of $n - r$ actors was constructed. In order to derive scale-values for the r actors, they could (a) be paired with each of the $n - r$ actors—thus making $r(n - r)^1$ comparisons, or (b) the r actors could be paired with one another *as well as* with the $n - r$ actors, thus making $r(r - 1)/2 + r(n - r)$ comparisons. Both of these procedures were carried out.

To summarize, then, one criterion scale and three sets of so-called "derived" scales were developed:

1. Criterion scale—all ten items used—$n(n - 1)/2 =$ forty-five comparisons.

2. Single-item insertion into a previously established scale of nine items—$n - 1 =$ nine comparisons. (Ten such scales.)

3. Multiple-item-insertion into a previously established scale, making only $r(n - r)$ new comparisons in each case. (Six such scales.)

4. Multiple-item insertion into a previously established scale of $n - r$ actors, making $r(r - 1)/2 + r(n - r)$ new comparisons in each case. (Six such scales.)

Results

The following tables and figures are presented for reference:

Table 1 is the original matrix of comparison-judgments showing the percentage of times the row-object-choice is preferred to the column-object-choice. Table 1 also gives the criterion scale, all values considered positive, which was derived from the matrix.

Table 2 presents the ten scales derived from the method of single-item insertion, all values positive. The correlations between each of the scales and the criterion scale appear at the foot of each scale.

---

[1] While the main diagonal of the percentage-preference matrix, corresponding to self-comparison of each item, was actually used in the construction of all these scales, it is not included in the discussion above.

Table 2

Single-Item Inserted Scales

Note: The column headings refer to the actor who was withdrawn and re-inserted into the scale of the remaining nine.

Actor	Criterion Scale	A	B	C	D	E	F	G	H	I	J
A	190	334	160	185	158	184	190	130	190	185	185
B	705	836	699	682	655	686	762	655	645	705	659
C	962	1072	966	929	966	948	978	844	904	969	921
D	986	1098	1020	969	926	996	1002	942	904	969	921
E	685	789	709	705	655	679	694	655	622	659	612
F	559	673	523	566	539	568	583	539	506	566	473
G	791	907	803	871	772	780	787	731	602	799	682
H	000	000	000	000	000	000	000	000	000	000	000
I	875	976	874	871	868	898	881	820	809	842	824
J	433	484	429	425	421	452	482	446	316	425	400
	r	.995	.998	.995	.998	.999	.998	.995	.995	.999	.996

Tables 3 and 4 similarly present the scales derived from the method of multiple-item insertion. Table 3 gives the scales for the $r(n - r)$ comparisons, and Table 4 gives the scales for the $n(n - 1)/2 + r(n - r)$ comparisons. Correlations between each of these scales and the criterion scale similarly appear at the foot of each scale.

Figures 1, 2, and 3 are presented to illustrate graphically the results shown in Tables 2, 3, and 4. Figure 1 is a graphical presentation of the appearance of each item on the scales of preference of actors. The

Table 3

Multiple-Item Inserted Scales
(Scale Value for Each Inserted Item Determined
on the Basis of $r(n - r)$ Comparisons)

Note: The column headings refer to the actors withdrawn and re-inserted.

Actor	Criterion Scale	F.	F.G.	F.G.H.	F.G.H.B.	F.G.H.B.D.	F.G.H.B.D.J.
A	190	190	121	118	058	000	000
B	705	762	711	640	625	546	475
C	962	978	844	759	775	774	606
D	986	1002	954	852	884	795	685
E	685	694	658	580	600	559	432
F	559	583	557	502	444	406	203
G	791	787	721	605	595	546	349
H	000	000	000	000	000	008	047
I	875	881	818	731	706	686	588
J	433	482	502	370	355	350	306
	r	.988	.988	.992	.988	.984	.943

Table 4

Multiple-Item Inserted Scales

(Scale Value for Each Inserted Item Determined

on the Basis of $r(n - r) + \dfrac{r(r - 1)}{2}$ Comparisons)

Note: The column headings refer to the actors withdrawn and re-inserted.

Actor	Criterion Scale	F.	F.G.	F.G.H.	F.G.H.B.	F.G.H.B.D.	F.G.H.B.D.J.
A	190	190	121	160	115	146	190
B	705	762	711	682	705	705	705
C	962	978	844	801	832	920	886
D	986	1002	954	894	941	986	986
E	685	694	658	622	657	705	622
F	550	583	517	559	559	559	550
G	791	787	749	791	791	791	791
H	000	000	000	000	000	000	000
I	875	881	818	773	763	832	778
J	433	482	502	412	412	490	433
	r	.998	.989	.990	.900	.095	.994
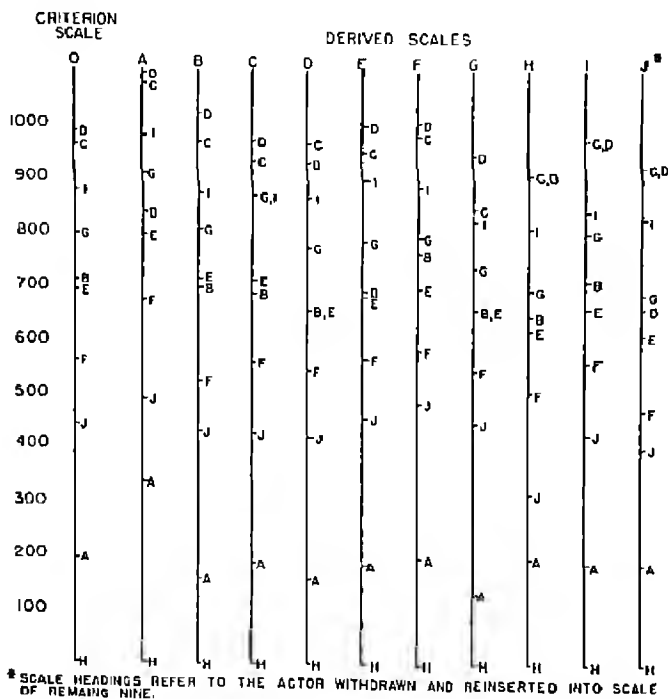
FIG. 1.   Single-item inserted scales (see Table 2).
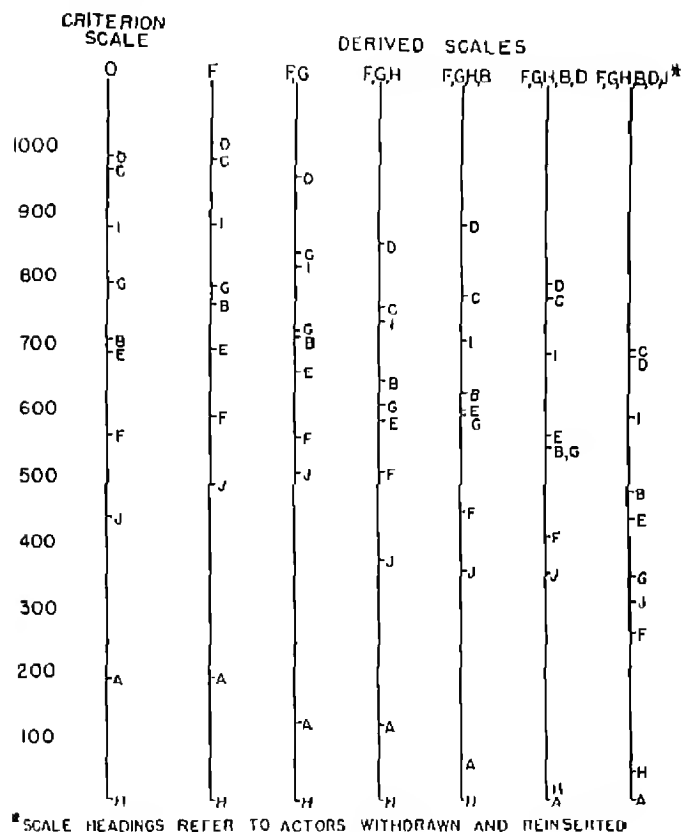
FIG. 2. Multiple-item inserted scales (see Table 3).

criterion scale is presented here along with the separate scales derived from single-item insertion. Figure 2 gives the scales for multiple-item insertion where $r(n - r)$ comparisons were made; and Figure 3 gives the multiple-item insertion scales when $r(r - 1)/2 + r(n - r)$ comparisons were made.

As may readily be seen from the tables and figures above, there is little doubt that substantially nothing has been altered in the construction of the "derived" type of scale. The scale resulting from inserting items into a pre-established scale differs negligibly from a scale developed with the use of all possible paired comparisons. The correlations between the "derived" scales and the criterion scale are, in every instance, .94 or over, even when the number of items inserted into the pre-established

*SCALE HEADINGS REFER TO ACTORS WITHDRAWN AND REINSERTED
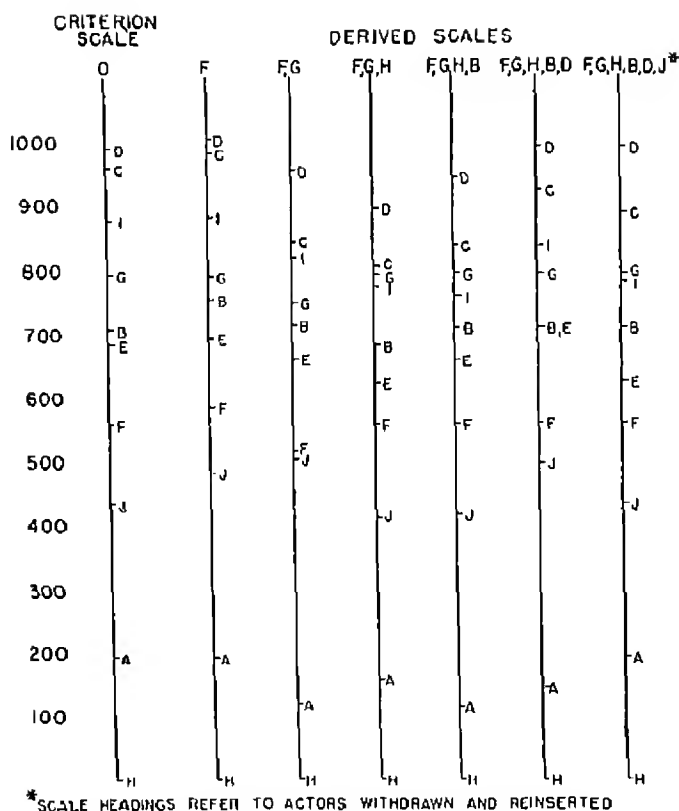
FIG. 3. Multiple-item inserted scales (see Table 4).

scale exceeds the number already in the scale. In all but one instance the correlations are .98 or greater.

Conclusions

Generally speaking, it appears that the smaller the number of items inserted, the higher will be the validity[2] of the "derived" scale. It is felt that the validity is roughly inversely proportional to the percentage of items inserted to the number in the pre-established scale. Particularly when the scale is not a reliable one—as is probably true in the present case—insertion of more than 50% will tend to lower the validity of the scale beyond desirable limits. In the opinion of the author, the ratio,

[2] In the usual sense of the term, "validity" does not strictly apply here. What is meant by "validity" is the correlation of a "derived" scale with the criterion scale.

$r/n - r$, should be no greater than .50. The implication for job evaluation is that when fifty per cent of the present jobs are altered, or a correspondingly similar proportion of new jobs is added, a new scale of n items should be drawn up. Even here, it should not be necessary to make $n(n - 1)/2$ new comparisons. It would be sufficient to retain the $(n - r)(n - r - 1)/2$ old comparisons, and to add to that $r(r - 1)/2 + r(n - r)$ new comparisons in order to build a new matrix and scale from the total of $n(n - 1)/2$ comparisons.

In general, the greater the number of judgments possible for the r items, the higher will be the validity of the "derived" scale. That is, when the r new items are compared one with another as well as with the $n - r$ old items, higher correlations result between the "derived" scale and the criterion scale.

It appears from this study that much can be done in the way of modifying the construction and use of the paired-comparison scale without altering appreciably the units along the scale. It is felt that such stability deserves further investigation of the paired-comparison technique. Unfortunately, as the number of object-choices increases, the number of paired judgments increases so rapidly that the scale falls down under its own weight. Additional work is needed, then, to test further the modifiability of the technique in order to permit a wider range of application.

The applications of this kind of modification in technique are fairly numerous. In the construction of attitude scales, for example, it has often been experienced by workers in the field that attitude statements, while meaningful during a particular period or for a particular group of subjects, lose their applicability and meaning with the passage of time or with a change in the characteristics of the group measured. It is at that time necessary to delete items from the original scale, and sometimes necessary to add new ones. It is apparent that any change in an item of the scale will change the complexion of the rest of the items in the scale. The question to be answered, then, is whether or not the scale resulting from the change in one or more items is sufficiently large to warrant an entirely new scaling of all items. To a considerable extent the present study answers such a question. If the empirical findings here continue to obtain, this type of manipulation with the items of a scale that has been derived by means of a paired-comparison technique can become quite extensive before the derived scale can be considered invalid.

The implications of this technique of "derived" scales for industry are fundamental from a more general point of view. Fortunately or unfortunately, the industrial situation seldom meets the rigorous as-

sumptions involved in statistical techniques that are developed on the
statistician's desk or in the laboratory. Particularly in the industrial
situation where personal satisfactions and benefits depend so heavily on
the assignment of a rating or judgment, is it important that the judgments
be made with greatest regard for precision and care. Ideally the pro-
cedures adopted for use should conform with the procedures found to be
most reliable in the laboratory. However, to the extent that practical
considerations make impossible the use of orthodox scientific techniques,
modifications must be introduced to conform with what is practicable.
Still, from the point of view of scientific awareness alone, if nothing else,
it is similarly necessary to know precisely what is the extent of reduction
in the validity and reliability of a measuring instrument or technique as
a result of modifying the orthodox procedures. It is only when he is
equipped with such knowledge that the psychologist can deal with his
data in industry with any real assurance.

References

1. Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1936.
2. Thurstone, L. L. The method of paired comparison for social values. *J. abnorm. soc. Psychol.,* 1027, 21, 384–400.
3. Uhrbrock, R. S., and Richardson, M. W. Item analysis. *Person. J.,* 1933, 12, 141–154.

The Paired Comparison Technique for Rating Performance of Industrial Employees

C. H. Lawshe, N. C. Kephart, and E. J. McCormick

Occupational Research Center, Purdue University

The method of paired comparisons has been used occasionally for making subjective ratings of job performance but has not been commonly adopted for this purpose, presumably because of certain disadvantages involved in its usual application. These disadvantages especially center around two factors: first, the time required, including the preparation of the pairs of names of the subjects, the actual rating process, and the summarizing of the results; and second, the rating process has been considered wearying to the raters, particularly if a considerable number of individuals are to be rated.

The Personnel Comparison System [1]

The *Personnel Comparison System* provides for the rating of job performance by the paired comparison technique, but the mechanics of its administration were specifically designed to simplify the various procedures.

The system lends itself to rating any aspect of employee performance, although in most of its applications it has been used for rating over-all job performance. The cue for the use of this basic factor as a measure of job performance is derived from such studies as that of Ewart, Seashore, and Tiffin (1) which brings out high degrees of communality among factors typically "measured" on rating scales. These authors identified the factor "Ability to do the present job" which accounted for most of the variability of the ratings.

The *Personnel Comparison System* provides the rater with a booklet composed of slips of paper about one inch wide and six inches long. See Figure 1. Each slip contains one pair of names. To facilitate preparation, eight slips are initially arranged on one $8\frac{1}{2}$ by 11 form. Pairs of names are typed on each slip and the slips are later separated by tearing along perforated lines.
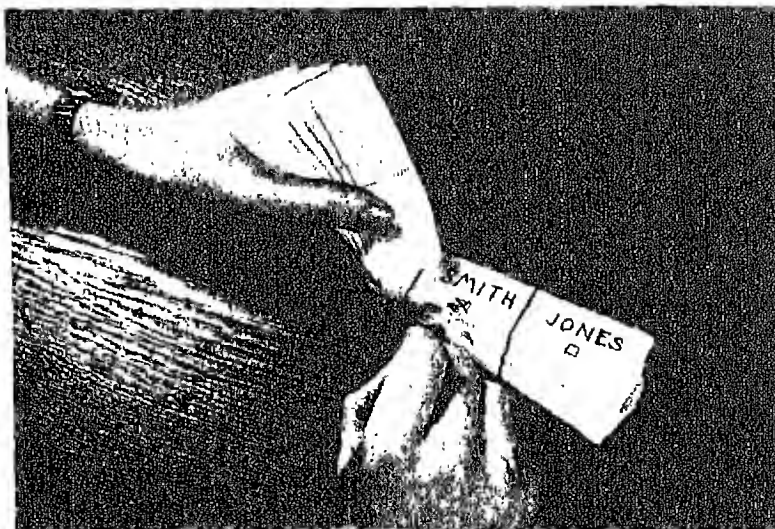
Fig. 1.   Method of marking pairs in the rating booklet made with the
Personnel Comparison System materials.

The procedures involved in the administration of the system and the
subsequent scoring of results follow:

1. The names of individual pairs are typed on the separate sections of
the forms according to a pre-determined order which is presented in table
form.   The table provides for pairing each employee with each other
employee.

2. The sections are separated on the perforations and the slips are
assembled into a booklet by means of a paper fastener inserted through
prepared holes.

3. The rater checks the preferred name on each slip.

4. The number of times each individual is preferred is tallied on a
summary sheet.

5. A performance rating index is derived from a table,[2] the specific
index being determined by the number of times each individual was
preferred and the number of individuals being rated.

[2] The indexes in the table are based on the proportion of times each individual is
preferred, converted to standard score units.   These units are based on a mean of 50
with a standard deviation of 10.   Indexes range from approximately three standard
deviations below the mean to approximately three standard deviations above the mean
(actually from 23 to 77).

Application of the System

For the purpose of experimentally applying the *Personnel Comparison System* in an operating situation, arrangements were made with a paper form manufacturing company to rate employee performance in two selected departments. This experimental tryout was directed toward the establishment of a criterion for the validation of personnel tests, rather than as a merit rating procedure. The raters were asked to rate the individuals with the following question in mind, "Which of these two employees is performing his present job better?" The two departments in which the system was tried, and the specific provisions for the application of the system in each, are given below.

1. *Offset press department.* Twenty-four of the offset pressmen who were oldest in point of service were rated by three supervisors and an instructor. All four raters had had an opportunity to become familiar with the work of each pressman through the systematic rotation of the pressman from one shift to another. One booklet including all pairs of employees was provided for each rater. Ratings were made independently.

2. *Stereo press department.* Eight stereo pressmen on each of three shifts were rated. These 24 men had had five or more years of experience on the job. While each man had previously been rotated between all three supervisors, they were classified in terms of their present shift position and the men on each shift were divided into random halves, called, 1–1, 1–2, 2–1, 2–2, 3–1, and 3–2.

On one day, each of the three supervisors (designated A, B, and C) rated those men then on his shift. On the next day, each supervisor rated the same men along with one-half of the men then on each of the other shifts. The groups rated by the three supervisors on the first and second days are indicated in Table 1. In addition, an instructor (Rater D) rated all of the 24 men.

Table 1

Subgroups Rated by Each of Three Supervisors on Two Days

Group	Supervisor Making Rating			
	First Day	Second Day		
1–1	A	A		C
1–2	A	A	B	
2–1	B	A	B	
2–2	B		B	C
3–1	C		B	C
3–2	C	A		C

Results of Offset Pressmen Study

The first study, involving the 24 offset pressmen, was conducted to determine the reliability of the ratings of different raters.

The agreement between the four raters is shown in Table 2. This table shows the number and per cent of pairs in which the same individual was preferred by all four raters; the number of pairs in which three raters chose the same man; and the number of pairs in which the raters split two-to-two. Of the 276 different pairs rated, all four raters preferred the same individual in 227, or 82.3 per cent, of the pairs. In 36 pairs, or 11.1 per cent, three raters preferred the same individual. In the 13 remaining pairs, or 4.7 per cent of the total, two of the raters preferred one of the individuals, and the other two raters preferred the other individual.

Table 2

Distribution of Preferences of Four Raters on Pairs of Twenty-four Offset Pressmen by Number and Per Cent of Pairs

Distribution of Preferences of Four Raters on Pairs of Employees	No. of Pairs	Per Cent of Pairs
4-0	227	82.3
3-1	36	13.0
2-2	13	4.7
	276	100.0

Intercorrelation of Ratings. Further analysis of the agreement among the four raters was accomplished by means of an average intercorrelation coefficient of the rank orders of the 24 men as resulting from the ratings of each of the four raters; the resulting average intercorrelation coefficient of the four rank orders was .94.

Reliability of Ratings on Halves and Quarters. In order to examine possible differences in reliability that would result from the rating of smaller groups of the same employees by the four raters, average intercorrelations were also computed for chance halves and chance quarters of these 24 offset pressmen. The two chance halves included odd-numbered and even-numbered employees respectively, the numbers having been assigned by alphabetical order of names. The chance quarters, in turn, were made up of every fourth name in the list in the same fashion. Only the preferences on pairs of employees included in the particular chance half or chance quarter in question were considered. Within each such group the number of times each employee was preferred was tallied, and

rank orders of the men in each group were subsequently determined. The average intercorrelations, computed by the rank-order method, are given in Table 3. These average intercorrelations closely approximate the coefficient of .94 obtained with the whole group. Even the correlation of .85 can reasonably be considered as satisfactory since only six men are involved.

Table 3

Average Intercorrelations of Rank Order of Times Preferred of Chance Halves and Chance Quarters of Twenty-four Offset Pressmen

Group	Average Intercorrelations
Chance halves	
1st half	.96
2nd half	.93
Average of 2 halves	.94
Chance quarters	
1st quarter	.97
2nd quarter	.85
3rd quarter	.93
4th quarter	.94
Average of 4 quarters	.92

Reliability of Ratings on Restricted Range Group. A further analysis of this same character was made with respect to a selected group of the 24 pressmen representing a restricted range of talent. The overall group included three floormen (working supervisors), thirteen "A" pressmen, seven "B" pressmen, and one helper. The 13 "A" pressmen (who operate somewhat more complex offset presses) were selected from the group for separate analysis, and the number of times each of these was preferred over the others within this same group was tallied. The resulting average intercorrelation of the rank orders of this group was .79.

This reduction in average intercorrelation from that of the overall group and from those for the chance halves and chance quarters would be expected since the group of "A" pressmen was much more restricted in its range of talent, and, generally speaking, tended to fall within the central and above-average (though not extreme top) range of the distribution of the entire group. The floormen consistently were rated above the "A" pressmen, and to a considerable extent the "B" pressmen and the helper tended to be rated toward the lower end of the over-all group.

The ratings of these 13 "A" pressmen were then subjected to a different type of analysis. The relative rank orders of these 13 men were "extracted" from the rank orders of the entire group; they were then

compared with the rank orders resulting from the preferences on *only* the pairs of men in this sub-group. The rho correlation between these two rank orders was .996, *indicating that there was practically no displacement* in rank-order position among these 13 men when their rank order was derived from the ratings made exclusively on this group, as compared with their relative rank orders when "extracted" from that of the whole group.

Results of Stereo Pressmen Study

As indicated before, the eight stereo pressmen on each shift were split into chance halves. On one day each supervisor rated the eight men together, and on the subsequent day each supervisor rated the same eight men along with one of the halves of each of the other shifts. The instructor rated all 24 men on one occasion.

In order to determine the correlations between subsequent ratings on men rated twice by the same supervisor, or on ratings by two or more raters on men rated in common, only the pairs of names pertinent to any such specific analysis were used in tallying the number of times each man was preferred. The rank-difference correlation coefficients (rho) between the several combinations of ratings are given in Table 4.

Table 4

Rank-Difference Correlations (Rho) of Various Ratings on
Twenty-four Stereo Pressmen

Rater	Groups Rated	No. of Men in Group	Coefficient of Correlation (Rho)
First and Second Ratings by Each of Three Supervisors			
A	1–1, 1–2	8	.98
B	2–1, 2–2	8	1.00
C	3–1, 3–2	8	.94
Average			.97
Ratings by Two Different Supervisors			
A & B	1–2, 2–1	8	.81
A & C	1–1, 3–2	8	.83
B & C	2–2, 3–1	8	.86
Average			.83
Ratings by Each of Three Supervisors and One Instructor			
A & D	1–1, 1–2 2–1, 3–2	16	.88
B & D	1–2, 2–1 2–2, 3–1	16	.90
C & D	1–1, 2–2 3–1, 3–2	16	.83
Average			.87

Reliability of Two Ratings by Three Supervisors. The initial analysis of the ratings of the stereo pressmen was that of the reliability of the two ratings made by each of the three supervisors of the eight men who were then under their respective supervision. The rank-difference correlations (rho) between the two ratings made by each of the supervisors ranged from .94 to 1.00, with an average of .97, which reflects a highly satisfactory degree of consistency between the ratings.

Reliability of Ratings Among Three Supervisors. As indicated above, eight men were rated in common by supervisors A and B, eight others were rated in common by supervisors A and C, and eight others were rated in common by supervisors B and C. The rank-difference correlations between the two ratings of each of these three groups ranged from .81 to .86, with an average of .83. While these coefficients between ratings made by different supervisors are somewhat below the coefficients of the two ratings made by the same supervisors on men whom they rated on successive days, they can nevertheless be considered as reflecting an adequate degree of consistency among the three raters.

Reliability of Ratings Between Three Individual Supervisors and One Instructor. Each supervisor rated 16 men, while all 24 were rated by the instructor. The rank-difference correlation coefficients between the ratings of each of the supervisors and the ratings of the instructor ranged from .83 to .90, with an average of .87.

Table 5

Average Intercorrelations of Ratings by Three Raters of Three
Groups of Eight Stereo Pressmen

Raters	Groups	Average Intercorrelation
A, B, D	1–2, 2–1	.84
A, C, D	1–1, 3–2	.76
B, C, D	2–2, 3–1	.87
Average		.82

Reliability of Ratings of Three Raters. Since each of three groups of eight stereo pressmen was rated by two different supervisors and by the instructor, it was possible to determine the average intercorrelations of the rank-orders resulting from the three ratings on each of these groups of eight men. These average intercorrelations were .76, .84 and .87, the average of the three being .82. (See Table 5.) This average is lower than the average of the other measures of reliability previously mentioned, but is within the same relative range as those of the other measures of reliability.

Administration of Rating System

Time Required for Administration. The time required for applying the rating system to the 24 offset pressmen may give a rough indication of the practical feasibility of the system in somewhat comparable circumstances. It was estimated that it took a total of 12 hours to type the slips for the 276 pairs (including carbon copies for the four raters), to assemble the four booklets, to rate the workers, and to derive the rating indexes. This time did not include planning, conference, or administrative time, but did include the time required for the rating by all four raters. In view of the fact that time required for functions such as typing and separation of the slips does not increase proportionately with the number of different raters, the over-all time is not indicative of the time that would be required if the rating were done by one rater rather than by four. It is estimated that the time required to prepare material and to summarize results for a complete rating of the 24 men by one rater would be about five or six hours.

The actual time required for each rater to rate the 276 pairs, however, was only about 30 minutes. This time required for actual rating is sufficiently reasonable to raise a question about the comments made by Guilford (2) and made in the report of the National Industrial Conference Board (3) to the effect that the method of paired comparisons is, by its nature, excessively wearying to the raters. More specifically, there is reason to doubt the limit of 15 subjects implied by Guilford as the upper limit of the practical application of the technique. Perhaps the mechanics of the specific scheme provided for making the ratings have a significant bearing on the degree to which the system is acceptable to the raters, and consequently on the total number of subjects that can reasonably be rated by one individual.

In considering the over-all time required for all the processes there was no suggestion that this time was considered excessive by the company applying the system to these two groups of workers.

Summary and Conclusions

Two groups of 24 workers each were rated by the paired comparison technique using the *Personnel Comparison System.* One of the groups included 24 offset pressmen who were all rated by three supervisors and one instructor. The other group included 24 stereo pressmen, eight from each of three shifts; each supervisor rated the eight men on his own shift on one day, and on the next day he rated the same men along with one-half of the men on each of the other shifts, making a total of 16 men. An instructor rated all 24 stereo pressmen once.

Analyses of the resulting ratings brought about the following primary conclusions:

1. There was a high degree of reliability between the ratings of two or more raters who rated the same employees.

2. There was a high degree of reliability between successive ratings, made on different days by each of three raters, on the employees whom they individually supervised.

3. The analysis of the ratings of a selected subgroup of employees revealed very little relative displacement in their rank-order position derived from the ratings on only the selected employees, as compared with their relative rank-order positions "extracted" from the ratings of the larger group of which they were a part.

4. The evidence accumulated did not indicate that the time required of raters was excessive.

References

1. Ewart, E., Seashore, S. E., and Tiffin, J. A factor analysis of an industrial merit rating scale. *J. appl. Psychol.*, 1941, 25, 481–486.
2. Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill Book Company, Inc., 1936.
3. *Employee rating; methods of appraising ability, efficiency, and potentialities.* National Industrial Conference Board, Studies in Personnel Policy No. 39, 1942.

Flesch Count and Readership of Articles in a Midwestern Farm Paper *

Howard B. Lyman

East Texas State Teachers College, Commerce, Texas

A preliminary study of the readers of *Wallaces' Farmer and Iowa Homestead* suggested in March, 1946 that reducing the Flesch count of articles from 3.5 to 1.5 might substantially increase the number of subscribers reading that article. To investigate this clue, a similar survey was set up in November, 1946.

The state of Iowa was divided into alternating counties, designated as "A" and "B" for the purposes of this report. The editor reveals that there may have been some sectional bias in the results, inasmuch as the "A" group of counties were a little heavier towards the southwest and the "B" counties heavier to the northwest.

Papers for November 16, 1946 were run off with four articles printed in alternate forms (one with a Flesch count of approximately 3.5, the other with a count of approximately 1.5[1], two of the difficult and two of the easy forms appearing in each copy of the issue. Typography, illustrations, leads, subject matter, and position of the articles were identical; only the difficulty level was varied. The experimental copy was distributed to all subscribers in the "A" counties, the control copy to all subscribers in the "B" counties. An excerpt from both forms of Article 4 (Nylons) is given in Figure 1.

Lower Flesch Count Version	Higher Flesch Count Version
Edna, my neighbor, was lucky. She has a big family. In 1940, she bought a pretty green nylon and wool coat for Bonnie, her eldest daughter. Bonnie wore the coat for two years. Then, when she became a war bride, she got a new coat that would match her wedding suit.	Nylon doesn't always mean just a precious pair of sheer stockings any more. It can mean any number of bright, new garments that are made of nylon. There are blouses, slips, children's clothes, coats and such things as curtains, rugs, and upholstery materials.

FIG. 1. Introductory paragraphs from Article 4 (Nylons).

[1] Flesch counts (1) computed by the United States Department of Agriculture Readability Unit.

From nine to thirteen days after publication, interviewers were sent out with instructions for obtaining a random sample of subscribers ("Go to first cross-roads north, turn to the right, call at every third farm"). If the farmer was a subscriber, he was asked: "Did you HAPPEN to see or read anything on this page?" Table 1 shows the final size of the sample after eliminating non-subscribers.

Table 1

Size of Subscriber Samples in "A" and "B" Counties

	"A" Counties	"B" Counties	Total
Male	73	76	149
Female	75	83	158
Total	148	159	307

A quantitative score for each article was obtained by the following system: a rating of 1 if the respondent indicated reading one-quarter or less of the article; 2, if one-quarter to one-half; 3, if one-half to three-quarters; and 4, if three-quarters to total. By using this system and correcting for difference in size of N and for variation in areas, it was possible to determine the per cent of greater readership for the articles with the lower Flesch count. These facts are presented in Table 2. The female scores for articles 1 and 3 and the male scores for article 4 were thrown out because of the small N's.

No figures on significance are reported, several statisticians having stated that the data are not amenable to any of the standard tests; however, after deriving the corrected quantitative readership score, it was found that four articles showed a positive difference (i.e., an increase in readers for the lower Flesch count) and one a negative difference. The low count version of the editorial (Article 2) showed a 9.4% decrease in readership for the women. Increases in readership for the other low count articles ranged from 7.3% to 66.0%.

Wallaces' Farmer reports that by deliberate attempts to keep copy more readable, they have been able to lower the range of most articles to between 1.5 and 4.0. Prior to this policy, articles had ranged from about 3.0 to 6.0. Routine reader surveys have shown consistent increases in readership, and the lower Flesch counts are considered at least a contributory factor to this increased popularity.

In the opinion of the writer, even better results may be anticipated from the use of the new Flesch readability yardstick (2). The old

Table 2

Difference in Readership Scores for Articles When Copy is
Simplified by Use of Flesch Principles

Article	Subject	County	Sex	Flesch Count	No. of Readers	Raw Readership* Score
1	Hogs	A	M	1.5	40	49.3
		B	M	3.85	50	57.0
		A	F	1.5	**	**
		B	F	3.85	**	**
2	Editorial	B	M	1.76	36	47.4
		A	M	4.27	24	32.5
		B	F	1.76	23	37.7
		A	F	4.27	23	30.7
3	Corn	B	M	1.35	51	65.8
		A	M	3.47	37	47.3
		B	F	1.35	**	**
		A	F	3.47	**	**
4	Nylons	A	M	1.11	**	**
		B	M	2.48	**	**
		A	F	1.11	44	52.3
		B	F	2.48	40	43.7

* After correction of the scores to make them comparable for the two groups, one low Flesch count article (2B Males) showed a loss of 9.4% in readership. Two others (1 and 3, Females) were dropped because of the small N. The other four showed increases ranging from 7.3% to 66.0%.

** Not computed, since N was less than 20.

formula yielded an ambiguous index in which difficulty and interest were combined. The new formula is not only simpler to apply but it measures difficulty and interest separately.

References

1. Flesch, R. *The art of plain talk.* New York: Harper and Brothers Publishers, 1946.
2. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221–233.
3. Murphy, D. R. Test proves short sentences and words get best readership. *Printer's Ink*, 1947, 218, 61–64.
4. Murphy, D. R. How plain talk increases readership 45% to 66%. *Printer's Ink*, 1947, 220, 35–37.

Speed of Reading Nine Point Type in Relation to Line Width and Leading *

Miles A. Tinker and Donald G. Paterson
The University of Minnesota

The writers have previously reported the optimal limits for good readability within which line width and leading may be varied for 6 point, 8 point, 10 point, 11 point, and 12 point type sizes.[1] The results appear to be specific for each type size. Nine point type was found to be as readable as 10 point, 11 point, and 12 point type when each was printed with two point leading, and with its own optimal line width.

It was believed to be important to establish for 9 point type the same information previously reported for each of the five type sizes mentioned above. The purpose of the present study, therefore, is to determine the influence of variation of line width and leading for 9 point type. The method used was the same as in previously reported studies.[1]

A table giving detailed tabulated results for the twenty test groups of 100 sophomore laboratory students each is on file with the American Documentation Institute.[2] This table is not reproduced here because of its excessive size and detail which would be of primary interest only to the research scholar.

Table 1, however, presents in convenient summary form a guide to be followed by those who desire to specify the optimal limits of variation in line width and leading when nine point type is to be used.

In setting up the study, we used, as a standard, material printed in 18 pica line widths with 2 point leading. The line width variations and the leading variations shown in Table 1 were each compared in turn with the standard. The differences are shown as percentage increases or decreases (minus sign) in speed of reading. For example, the test material printed in an 8 pica line width, set solid, was read 9.5 per cent more slowly than the standard, whereas the test material printed in the same short line width with 1 point leading was read 4.8 per cent more slowly than the standard. Other entires in Table 1 are to be interpreted in a similar manner.

[1] Paterson, D. G., and Tinker, M. A. *How to make type readable.* New York: Harper and Brothers, 1940. (Obtainable from the writers.) See Chapter 7, pp. 72–81. Also see Appendix I, Methodology, pp. 161–189.

[2] This table is available as ADI Documents in the form of microfilm (images one inch high) on standard 35 mm. motion picture film, or photoprints (6 × 8 inches in size) readable with unaided eyes. To secure this table order Document 2628 remitting $0.50 for photocopy or microfilm from American Documentation Institute, Science Service Building, 1719 N Street, N.W., Washington 6, D. C.

82 *Miles A. Tinker and Donald G. Paterson*

Table 1

Simultaneous Variation of Line Width and Leading for Nine Point Type

Note: Reading speeds for 8, 14, 18, 30 and 40 pica line widths each set solid and leaded 1 point, 2 points and 4 points are compared (percentage differences) with reading speed for Scotch Roman printed in 18 pica line width leaded 2 points as a standard. Minus (—) differences indicate slower reading than the standard. Figures in bold face indicate extremely unsatisfactory typographical arrangements. Number of readers — 2000 university sophomores.

Line Width	Set Solid	1 Point Leading	2 Point Leading	4 Point Leading
8	—9.52	—4.75	—5.76	—6.78
14	—4.39	0.68	0.46	1.30
18	—2.72	0.23	0.00	3.24
30	—5.17	—0.45	2.43	0.40
40	—5.83	—3.97	—5.81	—2.57

Examination of Table 1 shows that the region of optimal legibility ranges between a 14 pica line width with 1 point leading or more to a line width of about 30 picas with 1 point leading or more.

All differences amounting to a 4 per cent or greater *decrease* in legibility are indicated by the use of bold face type. Such differences are significant beyond the 1 per cent level. This permits ready identification of typographical arrangements that should not be used. The 2.72 per cent decrease in reading rate for 18 pica line width set solid is significant at about the 3 per cent level. While this arrangement may be used without a large retardation in reading rate, it is not recommended. The same is true of the 2.57 per cent decrease in reading rate for 40 pica line width with four point leading.

As was true with the other type sizes studied and previously reported, one can specify line widths for 9 point type over a considerable range (in this instance, from 14 to 30 picas) provided one to four points of leading are used. Conservative practice would probably specify one or two point leading for 9 point type in line widths varying from 16 picas to 24 picas. Our studies of reader preferences show that readers dislike long lines and very short lines.

Summary

The present study was carried out to determine the influence of line width and leading on the speed of reading 9 point type.

The results indicate that optimal rate of reading occurs with line widths of 14 to 30 picas and with 1 to 4 points leading. This may be considered the *zone of safety*.

A conservative range would be 16 to 24 pica line width with 1 or 2 points leading when 9 point type is used.

Received June 10, 1948.

Effect of Target Brightness on "Normal" and "Subnormal" Visual Acuity *

James E. Kuntz ** and Robert B. Sleight †

Division of Education and Applied Psychology, Purdue University

Ferree and Rand (1) and Ferree, Rand, and Lewis (2) have described the influence of illumination on the visual acuity of a few persons of widely divergent ages and with greatly varied visual abilities. They concluded (2): "Lighting practice has been conventionalized much too narrowly with respect to intensity of light." Tinker (3) believes that the studies referred to above ". . . suggest a moderate increase in illumination for those with corrected vision as compared with normal eyes."

The purpose of this experiment was to investigate this problem further by comparing the performance of a group of people with "subnormal" visual acuity with a group having "normal" visual acuity on a task of visual discrimination under varying brightness levels.

In this experiment those persons were considered subnormal who demonstrated a visual acuity *below* 1.0 and those considered normal who had a visual acuity *above* 1.0 in decimal notation when measurements were made at a distance of 28 inches and with a brightness level of ten footlamberts on the same test as used in the experiment reported on in this paper. There were 12 Ss in the subnormal group and 12 in the normal group.

Six brightness levels were used, viz., 3.16, 10, 31.6, 100, 316, and 1000 footlamberts. These conform in log terms to $10^{.5}$, 10^{1}, $10^{1.5}$, 10^{2}, $10^{2.5}$, and 10^{3}.

Apparatus and Acuity Targets

Figure 1 is a schematic diagram of the apparatus used in this experiment. Illumination was provided by electric lamps (L) varying in wattage from 15 to 500 mounted on a frame 52 inches square. The lamps were arranged so that glare in the visual field was almost entirely eliminated. In the center of the frame was an opening 6 inches square into which was fitted an eye shield (ES). The subject sat on seat (S) and read the target (T) at a distance of 28 inches through this opening. A light shield (LS) was provided to eliminate direct light from the lamps striking the S's eye. A hood (H) was also used to shield the subject from light reflected into the surrounding room.
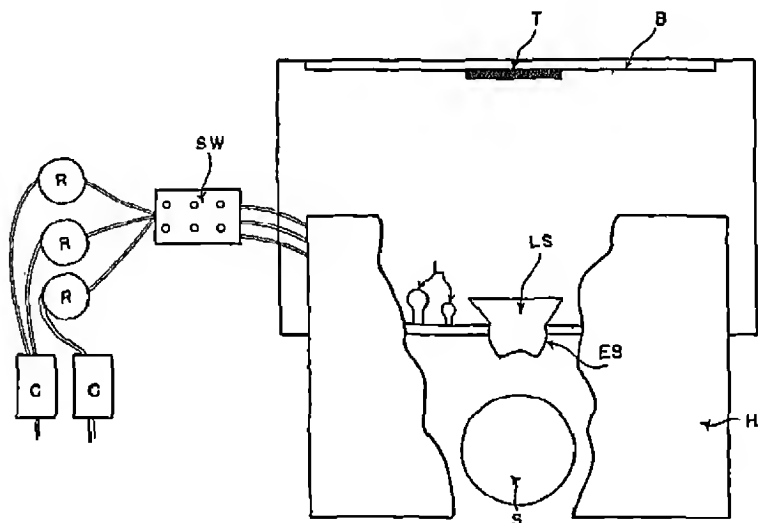
Fig. 1. Schematic diagram of apparatus used for the study of visual acuity (T = Acuity target, B = Background surface, LS = Light shield, ES = Eye shield, H = Hood, S = Seat, L = Lamps, SW = Switches, C = Constant voltage transformers, R = Rheostats).

The background (B) consisted of fine-grain wood covered with several coats of nonspecular white paint and extended to the limits of vision. The checkerboard target (T) was mounted on a piece of cardboard 6 inches square. The cardboard was of approximately the same texture and albedo as the background.

Constant voltage transformers (C) were used in the lines. The MacBeth Illuminometer was used daily, with measurements made at the position of the S's eye, to insure correct brightness levels. Brightness levels were readily changed by a series of switches (SW) which controlled the lamps of various wattage. Rheostats (R) were used to make very minor adjustments in brightnesses.

The acuity targets were of the type used in the Bausch and Lomb Ortho-Rater for measuring far and near acuity. Each target consisted of a checkerboard to be discriminated from gray areas in order to locate it in one of four

possible positions. The task consisted of locating the position of the checkerboard in a series of such targets progressively diminishing in size. The actual size of detail to be discriminated was as follows: 0.0135, 0.0102, 0.0080, 0.0067, 0.0058, and 0.0050 inches. (These sizes in terms of visual angle are 1.66, 1.25, 0.98, 0.82, 0.71, and 0.61 respectively.) The decimal acuity notations were determined by calculating the reciprocal of the visual angle subtended by the task object in each acuity target. The decimal notations which corresponded to each of the above targets were: .6, .8, 1.0, 1.2, 1.4, and 1.6 respectively.

Procedure

Targets were presented in a sequence designed to help eliminate the influence of different degrees of motivation, and to cancel the effects of learning and fatigue. The starting points (levels of brightness) for the Ss were rotated so that in each group two people began the experiment at each level and continued "up" the brightness scale until the highest brightness was reached. Those who did not begin at the lowest level then went to the lowest level and continued "up" to the point of beginning. There were 10 randomized presentations with regard to position for each target. If the S made five or more correct responses out of the ten target presentations, E then proceeded to administer the next smaller target until a target size was reached on which S made less than five correct responses in ten trials. All Ss were required to make a response for each target presentation. The succeeding target presentations were made in approximately 3 seconds after each response with the brightness constant at the level being used at the time. There was no time limit for making responses. Two minutes were allowed for adaptation in going "up" the brightness scale and 5 minutes when going from the highest to the lowest level.

The Ss used in this experiment were 7 female and 17 male students at Purdue University. The age range was from twenty to thirty-five years. One "normal" S was tested with glasses, the remaining Ss were tested without glasses. All Ss were tested monocularly, the unused eye being occluded by a card in the eye-shield.

Each S was instructed as follows: "This is a vision test. It is a rather complete test and will require approximately 40 minutes. You are to locate the position of the checkerboard in the target which is the same as the target used in the Ortho-Rater. (All Ss had been 'Ortho-rated' previously.) You are to respond with top, bottom, right, or left for each target presented. Each target will be presented 10 times. In order to determine how well you see, very small targets will have to be presented, so make the best response you can for each presentation even though you are not sure of the correctness of your response. Relax and rest your eye during the changes of brightness levels. You will be given a few minutes to get used to the next level of brightness. During that period keep your eye fixed on the target background."

Results

A decimal acuity score was calculated for each S by correcting the raw scores for chance. The following formula was used in obtaining this correction: $S_c = \frac{1}{3}(4C - 10)$ where S_c is the corrected score and C is the number of corrected responses.

The raw scores for each subject for a particular level of brightness were obtained by counting the number of correct responses on the target

James E. Kuntz and Robert B. Sleight

immediately following the smallest target on which at least 5 correct responses were made, then correcting this score by using the above formula. The decimal acuity score was obtained by interpolation. For example, S made a raw score of 4 on the fourth target in the series. The preceding target, no. 3, was the last target on which at least 5 correct responses were made. Target no. 3 subtends a visual angle of 1.0 giving a decimal acuity notation of 1.0 also. Interpolating for the interval of .20 (the difference in decimal acuity notation for targets no. 3 and no. 4) gave a decimal acuity notation of 1.040 for S for any one level of brightness.

The analysis of variance of the decimal acuities for the subnormal group is given in Table 1 and a similar analysis for the normal group is given in Table 2. The two analyses give essentially the same results

Table 1

Analysis of Variance of Decimal Acuities of the Subnormal Group

Source of Variation	Sum of Squares	df	Estimate of Variance	F
Between Brightness Levels	1.65	5	.329	32.90 *
Between Subjects	1.17	11	.107	10.70 *
Interaction	.56	55	.010	
Total	3.38	71		

* Significant at the 1% level of confidence.

Table 2

Analysis of Variance of Decimal Acuities of the Normal Group

Source of Variation	Sum of Squares	df	Estimate of Variance	F
Between Brightness Levels	.33	5	.066	4.71 *
Between Subjects	.18	11	.016	1.14
Interaction	.78	55	.014	
Total	1.29	71		

* Significant at the 1% level of confidence.

with the exception that the normal group could be regarded as more homogeneous than the subnormal group. The F values found show that levels of brightness play an important role in determining acuity scores for both groups.

Figure 2 shows graphically the relation between acuity and levels of brightness for the two groups. The mean decimal acuity for the subnormal group was .668 as compared to 1.057 for the normal group at the

lowest level of brightness, viz., 3.16 footlamberts.[1]   This level resulted
in the lowest acuity for both groups.   Maximum acuity was reached by
both groups at 1000 footlamberts, the mean decimal acuity being 1.061
for the subnormal group and 1.264 for the normal group.   The subnormal
group made a mean decimal acuity gain of .393 as compared to .203 for the
normal group.   The significance of the difference of mean gains resulted
in a "t" value of 2.54.   (A "t" value of 2.819 is required for 1% level of
confidence, 2.074 for 5% level of confidence.)

   The significance of the difference of the slopes of straight lines fitted
to the means of each group by the method of least squares resulted in
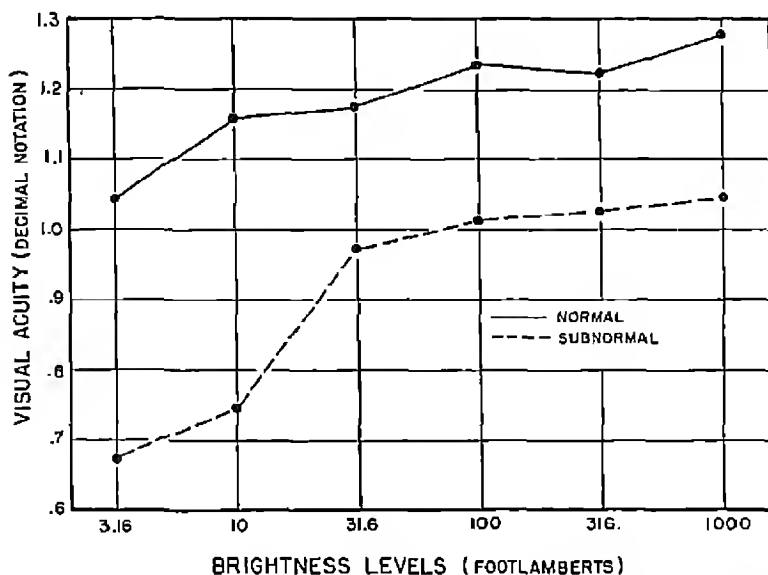
FIG. 2.   Variation of mean visual acuity for "Normal" and "Subnormal" Groups with
change in level of target brightness.   N for Normals = 12, N for Subnormals = 12.

   [1] The reader may be more familiar with light measurement in terms of footcandles.
The *footcandle* is a photometric measure which specifies the quantity of light falling
upon a surface.   A one-candle source delivers one footcandle of illumination on a
surface when the surface is at a distance of one foot.   The *footlambert* is also a photo-
metric measure, but it quantifies the amount of light coming back from a reflecting
surface.   A perfectly reflecting surface which has one footcandle of illumination on it
will have a brightness of one footlambert.   In order to measure the brightness of a
surface it is necessary to multiply the illumination on the surface (in footcandles) by
the overall reflectance of the surface.   For example, if a surface reflects 80% of the
light which falls on it then when one footcandle of illumination is put on the surface,
it will have a brightness of 0.8 footlamberts.   Apparent footcandle, another frequently
used brightness term, is equivalent to footlambert.

a "t" value of 3.15 which is significant well beyond the 1% level of confidence.   This technique, which takes into consideration all of the means of the two groups for each level of brightness, probably gives a more true indication of the effect of brightness increase than does a consideration of the two extreme means, viz., at 3.16 and 1000 footlamberts of brightness.

By further analysis it was determined that the lower one-half of the subnormal group, i.e., the six Ss showing lowest acuity, made a gain of .420 as compared to .393 for the entire subnormal group.   This gives some indication that the poorer the initial visual acuity the more beneficial an increase in target brightness becomes.

Also, as shown in Figure 2, if the curves were smoothed, it is of interest to note that the subnormal group reached "average" visual acuity (1.0 decimal notation) at a level of about 40 footlamberts with little change thereafter.   Further, it will be noticed that the normal and subnormal groups attained equal visual acuity at about 3.16 and 1000 footlamberts, respectively.

The per cent of maximum acuity is shown in Table 3.   The subnormal group benefited most from increased brightness as shown by a gain of 37.1% as compared to 26.4% for the normal group.

Table 3

Per Cent of Maximum Acuity * at Each Brightness Level

Level	Subnormal Group	Normal Group
3.16	62.9	83.6
10	70.5	91.6
31.6	91.5	92.9
100	95.4	97.5
316	97.5	97.1
1000	100.0	100.0
Total Gain	37.1	26.4

* Maximum acuity is defined as mean acuity of each group at 1000 footlamberts.

Tables 4 and 5 show the significance of the difference of means for the subnormal and normal groups between each level of brightness and every other level of brightness.

The only significant difference between means at successive levels of brightness is found between 10 and 31.6 footlamberts with the subnormal group, as shown in Table 4.   When comparing the mean at 3.16 with the means at each other level of brightness all are significant beyond the 1% level except one, viz., between means at 3.16 and 10.   It seems especially

Table 4
Critical Ratio of Differences Among Acuity Means at Each Level of Brightness for the Subnormal Group

Levels of Brightness	3.16	10	31.6	100	316	1000
3.16		1.95	7.40 *	8.40 *	8.94 *	9.59 *
10			5.44 *	6.45 *	6.99 *	7.64 *
31.6				1.01	1.55	2.10 †
100					0.54	1.19
316						0.65
1000						

* Significant at the 1% level of confidence.
† Significant at the 5% level of confidence.

Table 5
Critical Ratio of Differences Among Acuity Means at Each Level of Brightness for the Normal Group

Levels of Brightness	3.16	10	31.6	100	316	1000
3.16		2.06	2.40 †	3.61 *	3.50 *	4.25 *
10			0.34	1.55	1.44	2.20 †
31.6				1.21	1.10	1.85
100					−0.11	0.64
316						0.76
1000						

* Significant at the 1% level of confidence.
† Significant at the 5% level of confidence.

important to point out that there are no significant differences (at 1% confidence level) between brightness intensity 31.6 and any higher brightnesses within the range of brightnesses used.

As shown in Table 5 the critical ratios for the differences among means for the normal group are in general considerably smaller than for the subnormal group, although when comparing the mean at 3.16 with means at all other levels of brightness all differences are significant beyond the 5% level of confidence with one exception; the difference between the means at 3.16 and 10 is significant just slightly below the 5% level. One slight reversal can be noted, viz., between means at 100 and 316. With this group the fact that no significant gain (at 1% confidence level) in performance is obtained when the brightness is raised above 10 foot-lamberts may be of considerable consequence.

The average variability calculated by averaging the sigmas for each

group at each level of brightness was .154 and .103 (decimal notation) for the subnormal and normal groups, respectively. The standard deviation for the subnormal group for the lowest level of brightness was .150 and for the highest level of brightness .151. For the normal group the standard deviation at the lowest level was .175 and for the highest level .088.

Discussion

The findings of this experiment give additional confirmatory evidence that brightness is one of the primary factors in vision. Several previous investigations have shown conclusively that a person's visual acuity increases with target brightness. The present study, however, may be distinguished from most of the other investigations because it showed that the degree of gain in terms of ability to discriminate visually fine detail was relatively greater for those persons having initial below normal visual acuity, than for those having initial above normal acuity, when target brightness was increased.

It is not felt that the findings of this experiment warrant as specific a proposal concerning prescription of illumination intensities as was made by Ferree, Rand and Lewis (2): "In each case the individual needs should be determined and the intensity given that is required." However, it does permit the more generalized statement that when visual acuity is the primary concern, light levels should be relatively "high" on jobs requiring the seeing of details where persons with reduced visual acuities are employed. In all likelihood little advantage would be gained for these people by prescribing more than approximately 31.6 footlamberts. For individuals with normal visual acuity it is probable that there would be only slight advantage in prescribing more than approximatley 10 footlamberts.

It should be obvious why the authors of this article hesitate to recommend specific brightness levels for specific individuals. For one thing, the steps used were half-log steps and somewhat gross. Also, when individual cases are considered, certain complicating factors may be encountered even in evaluating a light level on the basis of threshold measurements. Not the least of these may be the motivational factors acting on the individual due to his relating a target brightness to the light under which he has been accustomed to working.

It should be borne in mind that this experiment was of a visual threshold nature. Tinker (3) believes that: "One should not prescribe illumination for suprathreshold tasks in terms of threshold measurements." The problem of optimum light intensities to minimize fatigue on prolonged tasks has not been covered in this investigation. Maximum acuity does not necessarily imply optimum working conditions because of

other variables which may be included in the overall situation. However, until a satisfactory criterion for "visual fatigue" has been ascertained it would seem desirable to utilize the findings from threshold experiments in choosing desirable light levels for "fine" tasks. Naturally specification of any feature of the working environment should take cognizance of the attitudinal viewpoint of the worker.

The findings of this experiment may have extensive ramifications, particularly in two interrelated endeavors, viz., (1) establishment of illumination standards, and (2) selection and placement policies in instances wherein an employee's visual acuity may be a factor in satisfactory performance.

Summary and Conclusion

An experiment was performed to determine whether the amount of increase in visual acuity, with increase of brightness on targets, differs *markedly for persons with initial "subnormal" acuity from those with* initial "normal" acuity. The experiment was of a threshold nature with subjects locating checkerboard targets under six levels of target brightness varying from 3.16 footlamberts to 1000 footlamberts.

It was found that a subnormal group gained significantly more in visual acuity terms with an increase in target brightness than did a normal group.

The data show that adequate light for seeing details is:[*]

(1) Between 10 and 30 footlamberts for those with normal vision;

(2) Somewhere between 30 and 40 footlamberts for those with subnormal vision.

References

1. Ferree, C. E., and Rand, Gertrude. The effect of intensity of illumination on the near point of vision and a comparison of the effect for presbyopic and non-presbyopic eyes. *Trans. Illum. Engng. Soc.*, 1933, 28, 590–611.
2. Ferree, C. E., Rand, Gertrude, and Lewis, E. F. The effect of increase of intensity of light on visual acuity of presbyopic and non-presbyopic eyes. *Trans. Illum. Engng. Soc.*, 1934, 29, 296–313.
3. Tinker, M. A. Illumination standards for effective and easy seeing. *Psychol. Bull.*, 1947, 44, 435–450.

Book Reviews

Lawshe, Jr., C. H. *Principles of personnel testing.* New York: McGraw-Hill Book Co., 1948. Pp. 227. $3.50.

This book is an elementary treatment of the problems involved in the testing of employees for purposes of selection. The expected phases are covered, namely test construction and validation (Chaps. II, III, IV and XIII), review of previous findings concerning the effectiveness of employment tests (Chaps. V through XII), and establishment of testing programs (principally Chap. XIV).

The approach taken by Lawshe will meet the approval of industrial psychologists. He is concerned with many of the problems in testing that can be appreciated only by one who has worked directly in industry. However, in the reviewer's opinion the book gives only a general overview of the field. The coverage of important problems, methods, and findings is spotty, and any reader, virgin to testing, may obtain an incorrect, and certainly an incomplete, picture.

The problems dealing with test construction and validation are adequate as far as they go—but they do not go far enough. The concept of reliability is not mentioned either in connection with criteria or with tests. There is no treatment of the combination or weighting of tests in a battery. Lawshe states that the purpose of the book is to serve as an aid to management. But the level at which the book is cast suggests that the author is underestimating the intellectual capacities of his potential readers. The point of view seems to be that concepts of second order difficulty, even though they be fundamental in nature, have no place in an introductory presentation. This simplified treatment is likely to convey to the wrong people the erroneous impression that anyone can develop and operate a testing program without getting into difficulties.

The chapters concerned with reporting previous findings on the validity of tests will be disappointing to many. It is by no means an extensive review. Rather these chapters are intended to present illustrative findings concerning the usefulness of different types of tests for various kinds of jobs. Several comments seem pertinent in this connection. Not a single example of the work of the U. S. Employment Service on aptitude testing is cited. Yet surely Stead and Shartle's now classic "Occupational Counseling Techniques" which summarizes so many of these excellent investigations deserves mention here. However,

of considerable value are the findings of a number of investigations not published elsewhere. Of the validation studies cited, only a very few with negative findings are given. Thus tests are hardly ever put in an unfavorable light. It reminds one of the optimistic descriptions of validity given in the manuals of directions of so many published tests. Any members of management or unions who believe this to be the true picture are in for a sorry disappointment. Even a cursory review of published reports will indicate that there is considerable variation in the effectiveness of any given test applied to different groups of workers on the same job.

Yet in fairness to the author it should be pointed out that he does not subscribe to the rather extreme views as presented in his publisher's advertising. Thus while the publisher claims that "From now on you can place the right person in the right job every time!", Lawshe emphasizes that tests are by no means a cure-all and can simply increase the probability of selecting better employees.

The third topic, establishment of testing programs, is likely to be dismissed as being another set of "practical" rules. This would be an error. While one might have hoped for a more expanded treatment, Lawshe here deals with most important problems. Today any book in this field purporting to be a "Principles" of employee testing would be woefully inadequate if it avoided such areas as supervisory support, budget, labor unions, personnel records and management reports. The day when the psychologist's task begins with the writing of items and ends with the computation of the validity coefficient is past, if it ever existed. Lawshe recognizes this and entertains for discussion certain of the important implications of testing in its larger setting of personnel problems and labor relations.

In sum, the reviewer's chief criticisms of this book are omissions of fundamental problems and concepts, and the elementary nature of the presentation. Those who use the book as a text either in college courses in testing or in similar courses for members of management or labor unions will undoubtedly find it necessary to provide supplementary material and discussion.

Edwin E. Ghiselli

University of California
Berkley, California

Chapin, F. Stuart. *Experimental designs in sociological research.* New York: Harper and Brothers, 1947. Pp. x+206. $3.00.

The breaking away of psychology from philosophy and the attempts of psychologists to convert their discipline into an exact science resulted

in a professional compartmentalization, the unfortunate effects of which
have only recently become clear to many psychologists.   In particular,
we have been insufficiently aware of the attempt of sociologists to estab-
lish their field as a science, for they broke away from philosophy after we
did and our natural science orientation has generally kept us from ob-
serving their efforts and progress.

Several works have in recent years attempted to review and con-
solidate the scientific gains made by sociologists: Lundberg's *Foundations
of Sociology* and Greenwood's *Experimental Sociology* come to mind as
illustrations.   Chapin's little volume is another distinctive contribution.
As he puts it in his preface, his purpose is "to illustrate the method of
experimental design by reproducing concrete studies," to provide "a
source book of examples of specific application (of the fundamental logic
of experimental designs) analyzed in some detail" (ix).   This is done by
analyzing the methods used in nine experimental studies.   Both the
methods and the findings are of interest to applied psychologists.

Chapin classifies the experimental designs used in sociological re-
search under three headings: the cross-sectional study, the projected
(before and after) design, and the ex-post-facto (retrospective) design.
His interest in these types of experimentation arises from the fact that
they can be used in real-life situations and are not limited to the labora-
tory or classroom.   He points out, for example, that social legislation
(e.g., slum clearance and work relief) is social experimentation, and his
interest is in experimental designs which can be used in the evaluation
of the effects of such experimentation.

The detailed analyses of experimental designs are stimulating in that
they point up the possibilities of research in practical situations, and
helpful in that they make clear the weaknesses and advantages of various
procedures.   Chapin brings out, for example, the inability of the social
scientist to emulate the natural scientist in controlling all but one of the
variables in an experiment.   He analyzes the alternatives and shows how
one of the best (randomization) is not generally feasible in real life (e.g.,
WPA workers were selected not only on the basis of need but also on the
basis of employability, thereby making them differ from the direct-relief
clients with whom they were to be compared in a study of the effects of
work relief on morale).   He then examines the use of available experi-
mental groups and control groups as a solution.   One such study (99-124)
evaluating the effect of high-school graduation on economic adjustment
(an ex-post-facto study) is analyzed to show the relative effects, on both
numbers and definitiveness of results, of the precise matching of indi-
viduals and of the grosser matching of distributions.   Matched distribu-
tions yielded experimental and control groups of 145 each from an original

group of 1194, contrasted with groups of 23 matched individuals each (matched for six variables). But the former method showed an insignificant difference in the economic adjustments of graduates and drop-outs, whereas the latter method showed that the graduates were clearly more successful. The emphasis on the need for the repetition of experiments in similar situations, in order to test the justifiability of generalizations, is also noteworthy.

A chapter and appendices dealing with sociometric scales should be of especial interest to psychologists. Some of these, both psychological and sociological, are already familiar, but the emphasis is on new instruments such as Chapin's Social Participation Scale and the revision of his Social Status Scale.

There is an interesting but, in this reviewer's opinion, unsuccessful attempt to justify cause-and-effect conclusions from indices of association. Chapin launches it by pointing out that the only alternatives are belief in chaos, magic, or means-end relationships. But the logic is fallacious, because the existence of such alternatives does not make it necessary to conclude that one of a particular pair of associated variables is the cause of the other. One may accept the principle of causation without being justified in concluding that, since the differences in the social adjustments of WPA workers and recipients of direct relief are statistically significant, and since the groups were matched on seven factors, work relief has a more beneficial effect than direct relief (p. 42). It is conceivable that better-adjusted relief clients were selected for work relief (the reviewer knows of WPA projects in which this was the standard practice), in which case superior adjustment was the cause of receiving work relief, rather than work relief being the cause of superior adjustment. Belief in causation does not indicate the direction of specific cause-and-effect-relationships. Statistics show association; the attribution of causal connections is a process of deduction. But this recurrent fallacy is of minor importance, provided the reader is aware of it.

This is a valuable book for those who are concerned with the design of experiments in social, clinical, educational, and vocational psychology, whether as research workers or as instructors. Its exposition is clear, it is rich in illustrative material, and the research principles which it illustrates are of widespread importance in the social sciences. It will also serve to introduce psychologists to an aspect of contemporary sociology of which many are too unaware.

Donald E. Super

Department of Guidance, Teachers College
Columbia University

J. G. Darley, Chairman, *et al.* *The use of tests in college.* Washington,
   D. C.: American Council on Education, 1947. Pp. vii+82. $1.00.
Froehlich, Clifford P., and Benson, Arthur L. *Guidance testing.* Chi-
   cago: Science Research Associates, Chicago, 1948. Pp. viii+104.
   $1.00.

Clarification of problems through fresh insights regarding them can
be a fruitful approach. Although this American Council on Education
publication is addressed to a college audience, it contains much of value
for the users of tests at most educational levels.

The frame of reference centers upon five questions, "Who shall be
admitted?", "How shall students choose appropriate curriculums?",
"How shall we counsel students?", "How shall we measure outcomes?",
"How do we measure behavior?" This method makes the material
useful to college and secondary school administrators, counselors, and
thoughtful instructors. The recommendations regarding the use of tests
are properly cautious and practicable.

A major value of the publication is its interpretative approach to test
use in terms of generalizations, rather than use of a multitude of specifics
related to particular tests. The reader is urged to consider carefully
Section V, *How shall we measure outcomes?* (pp. 43–57). The presentation
of an examination structure for colleges on pages 45 and 46 supplies an
excellent general framework for considering test results in relationship
to other types of data and several kinds of personnel workers.

In the reviewer's opinion, the objectives and implications stated in
the Foreword by Dean T. R. McConnell and Dean E. G. Williamson are
well met. It is his opinion also that it is a somewhat restricted but
generally excellent presentation. Strongly recommended reading for
student personnel workers and administrators in educational institutions.

Froehlich and Benson say: "This book is addressed to those individuals
who are faced with the responsibility of carrying on a guidance program
in which they must directly or indirectly administer and interpret tests,
even though their training in tests and measurements is limited (p. v)."

Guidance Testing poses again the problem of waiting until workers
are competent before using tests or urging that they gain this competence
by using tests cautiously. The book is definitely posited on the very
practical philosophy that naive personnel in education will use tests and
that it is sensible to help them to avoid errors in practice.

A major strength is the frank facing of the fact that proper use of
tests requires statistical knowledge so they include descriptions of simple
statistical methods and interpretations. This is in line with the view-
points of Bingham, Crawford, Darley, and others.

The authors present typical tests of various kinds under the rubrics

scholastic aptitude, achievement, interest, personal adjustment, and *special aptitude* (pp. 23–46). A footnote (p. 23) calls attention to the fact that the authors use these tests as examples, not as a selected list of the best instruments. Although agreement on the best instruments is probably impossible, some evaluation of excellence might have been helpful. One of the most frequently asked and most legitimate questions of the new test user is, "What are the best tests for my purposes?" Perhaps the answer is, "Consult the nearest competent person."

The feeling of the reviewer is that this is a helpful and useful book. This is particularly true if the audience for which it is intended follows through with graduate training which will make the book no longer needed.

Milton E. Hahn

University of California at Los Angeles

Erickson, Clifford E. (Editor). *A basic text for guidance workers.* New York: Prentice-Hall, Inc., 1947. Pp. 566. $4.25.

The editor of this volume, Dr. Erickson, who is Professor of Education and Director of the Institute of Counseling, Testing, and Guidance at Michigan State College, has written several previous books in the guidance field. Drawing heavily upon Michigan State College guidance instructors as well as upon a variety of guidance experts in other school systems, this text "attempts to portray many different aspects of the guidance program and at the same time to indicate the extent of some of the specializations within the field as a whole." In the preface the editor specifies that the book is intended as a basic or beginning text for training school counselors.

In general, the content of the book attempts to give the guidance worker (particularly the secondary school teacher) an over-view of the guidance field including purposes, techniques, and administration of the guidance program. Although the book fulfills its avowed aim to a large degree, the quality of the chapters varies widely and some overlap is evident as is often the case with such symposia. Aside from certain minor points the best chapters appear to be: the aims, objectives, and principles of the guidance movement (C. E. Erickson), interviewing techniques (S. A. Hamrin), therapeutic counseling (H. B. Pepinsky), helping pupils with their problems (P. L. Dressel), the community occupational survey (Elizabeth K. Wilson), the role of work experience (C. A. Weber), placement and follow-up services (L. O. Brockmann and L. Smith), and organizing the guidance program (C. M. Horn).

The major criticisms of the volume can be summarized briefly: unevenness in quality and treatment of material as well as in level of

difficulty (the chapter on therapeutic counseling may be heavy going for the average high school counselor); a tendency to apportion too much space to less important topics (case-study techniques and working with home and community); lack of critical evaluation of occupational information sources and of testing instruments; lack of functional information about the world of work in the job levels where most secondary school students will be employed.

The commendable features probably outweigh these criticisms, however. There is an admirable emphasis on the primary importance of individual counseling. The many good, practical suggestions are of great potential value to the guidance worker and should earn the authors many a word of praise from hard-pressed counselors. Another fine feature is the inclusion of many illustrative forms and excellent bibliographies.

Perhaps the most discouraging point raised by this book is the tremendous store of material, skills, and information the counselor must have as minimum working equipment. Insofar as guidance techniques can be put across in book form to the beginning counselor, *A Basic Text for Guidance Workers* is effective in attaining its aim.

<div style="text-align: right;">William A. McClelland</div>

Brown University,
   Providence, R. I.

Clarke, H. Harrison. *The application of measurement to health and physical education.* New York: Prentice-Hall Incorporated, 1945. $5.   p. 415.

This text is organized on a functional outline. After considering some of the fundamentals underlying testing in health and physical education, the measurement of physical fitness of social efficiency and physical education skills and appreciations are considered in turn.

In the section on physical fitness, there is the usual discussion of medical and sensory tests, of cardio-vascular tests, and a section on measurements and estimates of nutritional status. There is a well-written and sound treatment of the problems and possibilities associated with the technique of somatotyping. This is followed by an excellent discussion of measurement in the field of posture.

The author gives a somewhat undue amount of space to the so-called "physical fitness index," which is the percentage residual from the regression value of a general strength test. In the opinion of the present writer, far too much value is assigned to this test, and it is assigned attributes which it does not deserve. While strength is important in physical fitness, it does not deserve the reverence given here. The author again reverts to this test in the next part of his book.

In the part of the book given over to tests of social efficiency, the author has introduced a section that is new to testing textbooks in this field. He discusses—and gives numerous references to—a number of tests in the field of personality studies. This marks an advance in this field and these tests should be given more prominence in physical education and health studies as time goes on and as these tests improve.

In the field of skill tests, because of the large numbers of such tests, the author has had to choose a few of the more important ones to describe, and to give bibliographical references for the others. His choices have, on the whole, been good. The same has been true of his presentation of knowledge tests.

The text closes with a discussion of the administrative problems in physical and health education testing, and an appendix devoted to the not-uncommon attempt to compress a semester's course in statistical methods into a chapter.

On the whole, the book presents a number of fresh viewpoints and is a useful text and reference book in its field.

C. H. McCloy

The State University of Iowa

Ross, C. C. *Measurement in Today's Schools.* 2nd Ed.; New York: Prentice-Hall, Inc., 1947. Pp. xviii+597. $4.50.

Ross, C. C. *Chapter Exercises and Tests to Accompany Measurement in Today's Schools.* 2nd Ed., New York: Prentice-Hall, Inc., 1947. Pp. vi+74.

The second edition of this elementary textbook in educational measurement appears six years after the first edition. The organization, chapter and section headings, and approximately ninety-five per cent of the content remain unchanged. In general the references have been brought up to date and the content modified sufficiently to incorporate them. The chapter tests which appeared at the end of each chapter in the first edition have been supplemented by appropriate problems and placed in a consumable workbook to accompany the text.

The text with its accompanying exercise book is designed specifically for a first course in educational measurement. It is well organized to give the beginning teacher and partially trained school administrator an integrated insight into the theory of measurement, descriptive statistics and individual differences as they relate to school organization and instruction. Almost one-half of the text is devoted to the uses of measurement in motivation, learning, diagnosis, marking, grouping, promotion, guidance and evaluation. Appropriate emphasis is placed on the construction of informal teacher-made tests.

While the approach to the use of measurements is functional and integrative it is also traditional and uncritical. In the cataloguing of research the point of view is that of the professor of measurement rather than that of the director of school organization and learning. The extent to which our present knowledge of individual and trait differences in the schools points to new uses of measurement and needed reforms in school organization and practice is not sensed. Nevertheless these two books rate high among the teachable books available in the field.

Walter W. Cook

University of Minnesota

Erratum

In the December 1948 issue of the *Journal of Applied Psychology*, an error occurred in the article, "The Effectiveness of Intelligence Tests in the Selection of Workers" by E. E. Ghiselli and C. W. Brown. On page 576 the last eight lines of type should have been inserted following the first line on page 577.

Erratum

In the December 1948 issue of the *Journal of Applied Psychology*, under *New Books*, books by the following authors: H. L. Goldberg, K. Goldstein, D. T. V. Moore, Strauss and Lehtinen, L. R. Wolberg, and L. Szondi, were erroneously listed as being published by a book seller by the name of M. W. Drexler. The real publisher is New York: Grune and Stratton, Inc.

New Books, Monographs, and Pamphlets

Studies in psychosomatic medicine. Franz Alexander and Thomas M. French, Editors. New York: Ronald Press Co., 1948. Pp. 568. $7.50.

Some psychological apparatus: a classified bibliography. T. G. Andrews. Psychological Monographs No. 289. Washington, D. C.: American Psychological Association, 1948. Pp. 38.

Attitudes of German prisoners of war: a study of the dynamics of national-socialistic followership. H. L. Ansbacher. Psychological Monographs No. 288. Washington, D. C.: American Psychological Association, 1948. Pp. 42.

Foundation of psychology. Edwin G. Boring, Herbert S. Langfeld, and Harry P. Weld. New York: John Wiley and Sons, Inc., 1948. Pp. 632. $4.00.

Current trends in clinical psychology. A. W. Combs, et al. New York: The New York Academy of Sciences, 1948. Pp. 62.

The American woman in modern marriage. Sonya Ruth Das. New York: Philosophical Library, 1948. Pp. 185. $3.75.

Hearing and deafness. Hallowell Davis, Editor. New York: Murray Hill Books, Inc., 1948. Pp. 496. $5.00.

An application of the level of aspiration experiment to the study of personality. Sibylle K. Escalona. New York: Bureau of Publications, Teachers College, Columbia University, 1948. Pp. 132. $2.10.

The use of training films in department and specialty stores. Harry M. Hague. Boston: Harvard Business School, 1948. Pp. 147. $1.50.

Understandable psychiatry. Leland E. Hinsie. New York: The Macmillan Co., 1948. Pp. 359. $4.50.

Your job. Fritz Kaufmann. New York: Harper and Brothers, 1948. Pp. 238. $2.75.

A study of thumb- and finger-sucking in infants. Mary S. Kunst. Psychological Monographs No. 290. Washington, D. C.: American Psychological Association, 1948. Pp. 71.

Graduate training for educational personnel work. Corinne LaBarre. Washington, D. C.: American Council on Education, 1948. Pp. 54. $1.00.

The commonsense psychiatry of Dr. Adolf Meyer. Alfred Lief. New York: McGraw-Hill Book Co., Inc., 1948. Pp. 677. $6.50.

The strategy of job finding. George J. Lyons and Harmon C. Martin. New York: Prentice-Hall, Inc., 1948. Pp. 408. $3.25.

The open self. Charles Morris. New York: Prentice-Hall, Inc., 1948. Pp. 179. $3.00.

Educational psychology. Harvey A. Peterson. New York: The Macmillan Co., 1948. Pp. 550. $4.00.

Training employees and managers. Earl G. Planty, William S. McCord, and Carlos A. Efferson. New York: The Ronald Press Co., 1948. Pp. 278. $5.00.

The emotions. Jean-Paul Sartre. New York: The Philosophical Library, 1948. Pp. 97. $2.75.

The teacher as counselor. Donald J. Shank, et al. Washington, D. C.: American Council on Education, 1948. Pp. 48. $.75.

The legend of Henry Ford. Keith Sward. New York: Murray Hill Books, Inc., 1948. Pp. 550. $5.00.

Van Allyn methods manual. Keith Van Allyn. Palo Alto, Calif.: Surveys, Inc., 1948. Pp. 117. Manual plus 25 Qualification Inventories $7.50.

Cybernetics. Norbert Wiener. New York: John Wiley and Sons, Inc., 1948. Pp. 194. $3.00.

Pediatrics and the emotional needs of the child. Helen L. Witmer, Editor. New York: The Commonwealth Fund, 1948. Pp. 180. $1.50.

Diagrams of the unconscious. Werner Wolff. New York: Grune and Stratton, Inc., 1948. Pp. 423. $8.00.

Personnel management and industrial relations. Third Edition. Dale Yoder. New York: Prentice-Hall, Inc., 1948. Pp. 894. $5.00.

Exploring individual differences. Committee on Measurement and Guidance. Washington, D. C.: American Council on Education, 1948. Pp. 110. $1.50.

Exploring a first grade curriculum. New York Board of Education. Publication No. 30. New York: Bureau of Reference, Research and Statistics, Board of Education, 1947. Pp. 104. $.50.

Influencing and measuring employee attitudes. Personnel Series Number 113. New York: American Management Association, 1948. Pp. 55. $1.00.

Problems and experience under the labor-management relations act. Personnel Series Number 115. New York: American Management Association, 1948. Pp. 35. $.75.

New patterns of employee relations. Personnel Series Number 117. New York: American Management Association, 1948. Pp. 50. $1.00.

The Quantification of an Industrial Employee Survey.
I. Method *

Frank J. Harris †

Division of Education and Applied Psychology, Purdue University

The research project to be described is an attempt to develop a new technique to measure quantitatively the morale of industrial employees. In the past, two general approaches have been made to this problem. The first is an adaptation of the attitude scaling technique first described by Thurstone and Chave (3). By means of this technique a score is obtained which indicates the general attitude of employees toward the company for which they work. However, this type of scale does not provide management with very much insight regarding the attitudes of employees toward specific policies or practices.

The second approach consists of asking a number of questions about specific aspects of company policy. This type of employee opinion survey does provide management with the opportunity to obtain answers to those relatively specific questions with which it is often concerned. On the other hand, the answers determined from such a survey do not give overall attitude scores of the type required if departmental, tenure, sex, or other similar comparisons are to be made.

The present study is an extension of the general employee opinion survey approach. Briefly, the questions on the survey are statistically treated according to the generally accepted principles of test construction and standardization, thus combining the practical merits of the opinion survey with the quantitative aspects of the attitude scale.

We have been speaking of morale as if it were a term the definition of which was generally agreed upon. Such, of course, is not the case. However, the adequacy of this study will not stand or fall on terminologies

and for our purposes it has seemed unnecessary to go beyond an operational definition of morale. The definition adopted therefore is: "Morale is the attitude of the employee, as expressed on an anonymous questionnaire, toward the company for which he works, with a favorable attitude representing relatively high morale and an unfavorable or neutral attitude representing a relatively lower level of morale."

Development of the Morale Scale

Description of the original survey. The data were obtained from a survey conducted early in 1948 by the Victor Adding Machine Company in Chicago, Ill. The forms were mailed to the home addresses of all employees of the company. The forms were returned by the employees directly to Purdue University. The individual employees could not be identified in any way. The employees had been informed in advance of the nature of the project and their cooperation was requested in filling out and mailing the forms in an enclosed self-addressed, stamped envelope. Approximately 800 questionnaires representing 75% of the employees were returned. All of the data were coded and punched on I.B.M. cards for more convenient analysis. An analysis of the percentages of employees in various categories responding to each alternative was made quite independently of this study and forwarded to company officials.

Initial screening of items. Not all of the items in the original questionnaire could be presumed to be directly measuring attitude toward management or toward the company. Accordingly 48 questions were selected from the total which were considered to be appropriate to the study at hand. These 48 items with their alternative responses were then reproduced and presented to 10 judges with the following instructions:

The statements below are part of a questionnaire administered to employees of a manufacturing company. Kindly check the *one* response to each question which you think most strongly represents a favorable attitude toward the company.

The judges were all advanced students in or professors of industrial psychology. It was arbitrarily determined that items on which there was 80% agreement or better would be retained at this stage. Forty-six items met this criterion. In fact there was unanimous agreement on 42 items, 90% agreement on one, and 80% agreement on two. On one item of the questionnaire, which dealt with the filling of job vacancies, there were six possible responses. On this item, seven judges selected one response, while the three other judges selected a second response. It seemed logically justifiable to retain this item by considering either of these two alternatives as favorable.

Selection of stratified random samples. From the total number of questionnaires returned, those on which the respondent had failed to answer all of the biographical items were discarded. The remaining 753 were divided into two groups. These groups were randomly selected after the following stratifications had been imposed: male or female; married or single; weekly or hourly-paid; worker, set-up man or supervisor; length of service. One group, consisting of 377 employees, was considered the experimental group; the other, consisting of 376 employees, was held out for further analysis at a later stage.

Item analysis. A key card was prepared on which was punched the response to each item which represented "high-morale."[1] The cards of the experimental group were scored in terms of the total number of high-morale responses. The 100 highest scoring employees and the 100 lowest scoring employees were selected and the degree of internal consistency of each item was determined in terms of discrimination or D-values using Lawshe's nomograph (1). Items having a D-value of 1.0 or better were arbitrarily retained to comprise the scale. The 36 items which met this criterion, with their respective D-values, and with the high-morale response indicated, are presented in Table 1. It will be noted that, without any such intention on the part of the author, these items embrace many of the factors which various investigators have reported to be related to industrial morale. It is also worthy of mention that all of the final 36 items are those which the judges had previously agreed upon unanimously, provided either of two responses is accepted for item 36.

Reliability of the scale. At this point the experimental group of cards upon which the scale had been developed and analyzed was removed from further consideration. The second group which had been held out until this time was now scored in terms of the 36-item key. The odd-even reliability coefficient for this group was determined to be .72; correcting by means of the Spearman-Brown formula for the complete scale of 36 items yielded a reliability coefficient of .84.

Analysis of Morale Scores

Once the morale scale had been developed and was found to have satisfactory reliability, it was possible to proceed with an analysis of the scores of various categories of employees. The results of this analysis are shown in Table 2. In all of the comparisons presented the significance

---

[1] In this study, responses chosen by the judges as representing the most favorable attitude toward the company are termed "high-morale" responses. As used here the term may be considered as equivalent statistically to the term "correct" as it is customarily employed in item analyses of test items.

of differences was determined for group means and for group standard deviations. The significance of mean differences is expressed in terms of Fisher's t statistic; the significance of standard deviation differences is expressed in terms of Fisher's F-ratio as tabled by Snedecor (2). A t value which is significant at the 10% level of confidence is indicated by an asterisk. A t or F value which is significant at the 5% level is

Table 1

Final Morale Scale Items and Discrimination Values

Item		D-value
What Is Your Opinion of Your Boss (the Man You Report to)		
1. Does he "know his stuff"?	Yes..x..No......	1.20
2. Does he play favorites?	Yes......No..x..	1.35
3. Does he keep his promises?	Yes..x..No......	1.50
4. Does he pass the buck?	Yes......No..x..	1.10
5. Does he welcome suggestions?	Yes..x..No......	1.50
6. Is he a good teacher?	Yes..x..No......	1.70
7. Do the workers know more than he does?	Yes......No..x..	1.50
8. Does he set a good example?	Yes..x..No......	1.50
Do You Feel You Understand the Following Provisions of the Employees' Security Fund?		
9. How the money is divided among the employees?	Yes..x..No......	1.75
10. How the Company decides how much goes to this fund?	Yes..x..No......	1.80
11. How the Security Fund money is invested?	Yes..x..No......	2.80
12. How much you get if you leave, die or retire?	Yes..x..No......	1.70
13. Do you feel that you are receiving considerate treatment here?	Yes..x..No......	1.00
14. Do you feel top management is interested in the employees?	Yes..x..No......	1.90
15. Have you ever recommended this Company as a place to work to a friend?	Yes..x..No......	1.25
16. Do you feel you have a good future with this Company?	Yes..x..No......	1.85
17. What do you think of working conditions here as compared with other plants?		
	Above average..x..Average......Below average.....	1.60
18. How do you think your average weekly earnings (gross earnings before deductions) compare with that paid in other companies for the same type of work?	Better here..x..About the same......Lower here......	1.20

Give Careful Thought to the Following List of Company Policies Affecting Employees, Working Conditions, and Employee Benefits. Then Check What You Think About Each Item as It Is Being Carried Out.

Table 1 (Continued)

Item	Like	Dislike	Not Interested	D-value
19. Group Insurance Plan	..x..	....	.....	1.20
20. Security Fund–Profit Sharing	..x..	....	.....	1.50
21. Service Pin Awards	..x..	....	......	1.40
22. Vacations	..x..	....	......	1.50
23. Credit Union	..x..	....	......	1.30
24. Chance for promotion	..x..	....	......	1.60
25. Medical Department	..x..	....	......	1.00
26. Cafeteria	..x..	....	......	1.10
27. Lockers	..x..	....	......	1.00
28. Suggestion System	..x..	....	......	1.90
29. Employee Committees	..x..	....	......	1.50

30. Do you find your fellow workers:
         Friendly..x..Unfriendly......Indifferent......   1.20

31. What does your family think of this Company?
         Good place to work..x..No opinion......Poor place to work......   1.40

32. How do you like your present job?
         Very much..x..   Not so good......
         Pretty good......   Don't like it......   2.00

33. Do you think the employees have confidence in the operating heads of the business?
         Most employees do..x..   More than half of them......
         About half......Less than half......Few of them......   1.30

34. How do you feel your opportunities in this Company compare with those with your last employer?
         Better..x..Not so good......About the same......
         Never worked elsewhere......   1.70

35. What are your work plans for the future?
         Hope to remain here..x..Plan to work only a short time......
         Do not plan to work......I have other work plans......   1.35

36. When desirable job vacancies arise, how do you feel they are generally filled?

By both ability and service	..x..
By employing people outside the Company	......
By promoting favored employees who are not especially qualified	......
By giving first chance to employees of long service	......
By taking the most qualified person	..x..
I am not sure how they are filled	...... 1.30

indicated by two asterisks; at the 1% level by three asterisks. These asterisks are inserted for the convenience of the reader.

    Sex differences. It will be noted that although the mean morale scores

of the sexes do not differ significantly, the men are significantly more variable than the women.

Marital status. All married employees combined yield a significantly higher mean morale score than all single employees combined. In an attempt to determine whether either sex might account for this difference, a further breakdown was made. It may be seen that differences between the married and the single may be attributed primarily to the significantly higher scores obtained by married *men* as compared with single men.

Table 2
Comparisons of Morale Scores Among Employees Sub-Groups

	N	Mean	S.D.		t	F
Sex						
Male	253	26.99	6.41		.13	1.36**
Female	123	26.91	5.49			
Marital status						
Married	244	27.59	6.11		2.75***	1.04
Single	132	25.80	5.99			
Married men	178	27.66	6.28	MM vs SM	2.56***	1.06
Single men	75	25.39	6.47	MW vs SW	.97	1.20
Married women	66	27.39	5.69	MM vs MW	.32	1.21
Single women	57	26.35	5.20	SM vs SW	.94	1.55**
Type of job						
Worker	241	26.83	6.01	W vs S	.17	1.17
Supervisor	101	26.96	6.51	W vs S-U	1.02	1.12
Set-up man	34	27.91	5.67	S vs S-U	.80	1.32
Method of pay						
Weekly	98	27.76	6.26		1.68*	1.07
Hourly	278	26.48	6.48			
Weekly-paid worker	53	26.79	7.02		.05	1.16
Hourly-paid worker	188	26.84	5.70			
Weekly-paid supervisor	43	28.77	5.17		1.79*	1.86**
Hourly-paid supervisor	58	25.62	7.05			
Length of service						
Under 6 months	50	25.10	6.95		2.22**	1.64**
6 mos. to 1 year	86	27.66	5.42		1.66*	1.09
1 to 2 years	87	26.25	5.67		1.32	1.32
2 to 5 years	65	27.60	6.52		1.21	1.26
5 to 10 years	62	26.26	5.80		3.14***	1.05
Over 10 years	26	30.65	5.94			

* Significant at the 10% level.
** Significant at the 5% level.
*** Significant at the 1% level.

Marital status does not appear to affect the morale scores of women employees significantly. It is also evident that the greater homogeneity of women's scores is due more to the single than to the married women.

Differences in type of job. Employees were asked to indicate whether their job was best classified as that of a worker, supervisor, or set-up man. Comparative morale scores were determined for these three general categories. None of the differences between these groups is significantly greater than chance alone could reasonably explain, contrary to what one might expect on the basis of previously reported findings.

Weekly vs. hourly-paid jobs. The scores of all employees who were on a weekly salary were compared with the scores of all employees who were paid on an hourly rate basis. Since there was a tendency for weekly-paid employees to score higher than hourly-paid employees, further analyses of the data were made to determine whether workers or supervisors might account for this trend.[2] The results of this analysis indicate that the weekly-paid supervisors account for the higher morale scores of weekly-paid employees in general. Also, as a group, the scores of weekly-paid supervisors are more homogeneous than the scores of any comparable group.

Length of service. Employees were asked to indicate whether they had worked for the company (1) under six months, (2) from six months to one year, (3) from one to two years, (4) from two to five years, (5) from five to 10 years, or (6) over 10 years. Scores were analyzed in terms of these six categories. The results indicate that morale scores are lowest and most heterogeneous under six months. From six months to 10 years they appear to fluctuate to an insignificant extent. After 10 years they again take a significant swing upwards.

Summary and Conclusions

An attempt was made to develop a quantitative morale scale by treating responses to an industrial employee survey according to standard test development procedures. A questionnaire containing specific items of interest to management was filled out anonymously by approximately 75% of the employees of a Midwestern manufacturing company and mailed directly by each employee to Purdue University.

Questions which were obviously related to morale were judged by 10 competent individuals in terms of the one alternative response which represented a favorable attitude toward the company. The 46 items upon which there was 80% or higher agreement constituted the original scale.

[2] Set-up men were not included in this analysis since 32 of the 34 employees in this category were weekly-paid.

The questionnaires were then separated into two stratified random samples containing 377 and 376 cases respectively. One of these samples was scored and a high and low group of 100 cases each, based on total score, were selected. The per cent of the high scoring group and the per cent of the low scoring group responding to each item was determined. From these percentages the discrimination value of each item was computed. The 36 items having D-values of 1.0 or higher constituted the final scale. The sample which had been held out was scored in terms of the 36-item key. A corrected reliability coefficient of .84 was obtained by the split-half (odd-even) method.

The results of an analysis of the morale scores of various employee sub-groups are presented.

It should be pointed out and emphasized that the results of the analyses reported here are specific to the data from the particular company which cooperated in the study. Any attempt to generalize from these results as to the relative levels of morale among various groups of industrial employees would be hazardous.

No attempt has been made to explain the differences or lack of differences found. Such differences can be most safely interpreted by individuals thoroughly familiar with the plant from which the data were obtained. The results of the survey give such individuals a clue as to the focal points of the industrial relations program which might possibly call for special attention.

The methodology used in developing the scale may, on the other hand, be profitably applied to any industrial situation where data of the type described are available. The advantages accruing to any given company from this approach are several. In addition to the types of information usually obtained from an employee survey of this sort it becomes possible to:

1. Obtain a reliable and quantitative estimate of the relative morale levels of various groups of employees such as workers and supervisors, old and new employees, married and single employees.

2. Obtain a reliable indication of those areas in which a change in policy would seem desirable.

3. Secure comparable data with which to compare the state of morale from time to time and thus reflect the effect of any changes introduced by management.

4. Accomplish the above with little more effort than is involved in the treatment of the ordinary attitude survey.

Much more could be accomplished by further extensions of this approach. Working cooperatively through a common consultant a

number of companies would be able to obtain an indication of the level of morale of their employees as compared with employees of other companies. If it were possible also to relate the morale score of the worker to the supervisor under whom he works, industry would be better able to deal with one of the major sources of differences in employee attitude. It is hoped that the technique reported here will lead to further investigations along these lines.

References

1. Lawshe, C. H., Jr.   A nomograph for estimating the validity of test items.   *J. appl. Psychol.*, 1942, 26, 846–849.
2. Snedecor, G. W.   *Statistical methods.*   Ames, Iowa: Iowa State College Press, 1946. 485 pp.
3. Thurstone, L. L., and Chave, E. J.   *The measurement of attitude.*   University of Chicago Press, 1929.

The Quantification of an Industrial Employee Survey.
II. Application *

Frank J. Harris †

Division of Education and Applied Psychology, Purdue University

In a previous paper,[1] the author described a technique for developing a quantitative morale score by applying the principles and methods of test construction to an industrial employee survey. Advantages claimed for this approach are that the employer can secure comparable data with which to compare the state of morale from time to time, can obtain a reliable indication of those specific areas in which a change of policy might seem desirable, and is provided with a measure of the effect of any changes that are instituted. The present paper attempts to illustrate these advantages.

The survey from which the morale scale was developed was conducted in 1948. A similar survey had been conducted in 1945 for the same company, by the same consultant, and in the same manner. Of the 36 items selected from the 1948 survey to constitute the morale scale, 35 had appeared on the 1945 survey with minor modifications in wording in a few instances. In the earlier survey, 555 or 65% of the employees returned the questionnaire. Of the total respondents 60% were men and 40% were women. In the later survey, 800 or 75% of the employees responded of whom 66% were men and 34% women. Thus the two groups can be expected to be reasonably comparable in sex ratio and employee representation.

The results of the two surveys were examined to determine the direction and extent of any changes which might have occurred in the intervening three year period. A comparative study of this sort could be made in at least two general ways, depending upon the type of information desired. One way would be to score both sets of questionnaires in terms of the 36-item key. From the obtained scores it would be possi-

[1] Harris, F. J. The quantification of an industrial employee survey. I. Method. *J. appl. Psychol.*, 1949, 33, 103–111.

ble to compare the morale of employee sub-groups at the earlier and at the later time. Another way would be to compare the responses to each item by determining the per cent of employees who responded favorably to the item at each administration of the questionnaire form. The latter type of analysis was made in this study; for each item the difference was determined between the per cent of respondents who indicated a favorable response in 1945 and the per cent who indicated a favorable response in 1948. The level of significance of the differences between these percentages was determined by the computation of t-values.

The principal findings are summarized as follows:[2]

1. On 19 of the 35 items there was a shift in the high-morale or favorable direction at the 1% level of significance or better.

2. There was a favorable shift on three items at the 2% to 10% level of significance.

3. Only one item, "Does your boss play favorites?", showed an unfavorable change in attitude (at the 4% level of significance).

4. The remaining 12 items revealed slight changes in either direction which could be explained readily on the basis of chance alone.

In addition to changes in attitudes toward certain items an estimate was obtained of the general level of attitude toward each item or policy represented thereby. For example, although there was a markedly favorable *shift* in attitude toward comparative weekly earnings (from 12% responding favorably in 1945 to 37% responding favorably in 1948) the "level" of attitude remained rather low. On the other hand, 90% of the employees liked the group insurance plan in 1945 and in 1948.

The final interpretation of the findings and the uses to be made of them rest on decisions of the sponsoring company. The changes in morale or attitude were undoubtedly influenced to some extent by factors external to the company, e.g. conversion from war to peacetime production. However, management now has reliable indices of how certain of its practices have been received by the employees, has some definite clues as to what effects its policy changes have had on morale, and is in a better position to chart its future course in personnel relations.

"Item Analysis" Versus "Scale Analysis" *

Philip H. Kriedt and Kenneth E. Clark

University of Minnesota

In the last few years Dr. Louis Guttman of Cornell University has developed a new and increasingly popular technique for determining whether or not a test or attitude scale possesses unidimensionality (8). This paper presents a comparison of this technique of scale analysis with two older methods of item analysis, in order to determine the comparative values of each method for selecting from a pool of items those which belong together, either in terms of their internal consistency, or in terms of their unidimensionality.

The three methods herein compared are: (1) the Cornell Technique of Scale Analysis, in which the essential statistic is reproducibility, and in which emphasis is placed on the ability to predict or "reproduce" the response of an individual to every item of a scale in terms of his total score on that scale; (2) one common form of item analysis, specifically, that in which the item-responses made by persons in the top twenty-seven per cent of the distribution on total score are compared with the responses made by persons in the bottom twenty-seven per cent on total score, using the phi coefficient as a measure of the correlation between item and total score; and (3) the determination of inter-correlations between items as a means of selecting those which are measuring the same thing, using as the measure of relationship the tetrachoric correlation coefficient.

A 72-item Likert-type questionnaire on attitudes toward Negroes, made up of items with 3, 5, and 7 categories of response, was administered to 183 students in an elementary course in Social Science at the University of Minnesota. In general, the content of these items was extremely heterogeneous. The scale was scored initially by assigning arbitrary unit values to each of the response categories, as in the usual Likert-type scale.

Since the analytic methods being compared would be affected considerably by the methods used to reduce all item responses to dichotomies, some preliminary work was done to determine how best to group item responses. Response categories were first combined so as to maximize the per cent reproducibility of each item and, whenever possible, to make

each category have less error than non-error, in accordance with the requirements of the Scale Analysis methods. However, this approach did not seem particularly promising for the development of a good scale, since dichotomizing items so as to maximize reproducibility, without regard for other item characteristics, tends to provide dichotomies with high modal response frequencies; that is, items which are answered the same way by a large proportion of the respondents. Items were also dichotomized, therefore, on the basis of item correlation with total score. The top 27 per cent and the bottom 27 per cent on total score were selected, and their responses to every possible dichotomy of response categories were compared, using phi coefficients computed using Jurgensen's tables (9). That combination which maximized the phi coefficient was used. These dichotomies had few high modal response frequencies, since the method used tends to penalize items deviating markedly from a 50–50 split. Ten items were found to have such low phi coefficients for any combination of responses that they were not included in the later analyses.

For the remaining 62 items, all of which were now dichotomized, the following computations were made: all inter-item correlations, using the Cheshire, Saffir, and Thurstone computing diagrams (1) (method A); phi coefficients for each item versus total score on the 62 item dichotomized scale (method B); and the per cent reproducibility of each item as a part of the 62-item scale (method C). In addition, all seventy-two items were carefully read by the writers, and twenty-seven items selected as representing, in the judgment of the writers, the primary factor being measured by the scale. All questionnaires were rescored for these twenty-seven items, and reproducibilities computed for each of the items using this new total score (method D).

Four separate bases thus existed for the selection of items for a shorter, more unified scale. Using each method, a ten-item scale was constructed. These four 10-item scales will be referred to hereafter as scales A, B, C, and D. Scale A was made by selecting the ten items whose intercorrelations with each other would be maximized; scale B was made up of items with the highest correlation with the total score; scale C consisted of the 10 items having highest reproducibility in the 62-item computation; and scale D the 10 items having highest reproducibility selected from the special group of 27 items. The items in the two "reproducibility" scales (scales C and D) had no more error than non-error in each category, as required by the Guttman method. These two scales were almost identical, having eight of their ten items in common. None of the other scales, however, had more than three items in common.

A statistical description of each of these four scales is presented in Table 1. For each scale is reported: (1) reproducibility when rescored

Table 1

Item and Scale Characteristics of Ten Best Items on: Inter-Item Correlations (Scale A); Correlation Between Item and Total Score (Scale B); Reproducibility Using Pool of 62 Items (Scale C); and Reproducibility Using Pool of 27 Items (Scale D)

Distributions of % Reproducibilities

	A	B	C	D
95–99	1		2	2
90–94	3	2	6	2
85–89	4	3	2	2
80–84	2	4		
75–79		1		
N	10	10	10	10
Mean	88	85	98	94

Distributions of Phi Coefficients (Item vs. Total Score)

	A	B	C	D
90–99	3	3		2
80–89	4	3	2	
70–79	1	1	1	1
60–69	1	2	1	2
50–59		1	3	2
40–49	1		1	1
30–39				
20–29			2	2
10–19				
N	10	10	10	10
Mean	80	79	58	59

Distributions of Median Inter-Item r's

	A	B	C	D
80–84	4			1
75–79	5		3	4
70–74				2
65–69	1	1	1	2
60–64		4	4	1
55–59		1	1	
50–54		3	1	
45–49				
40–44		1		
N	10	10	10	10
Mean	73	57	65	72

Distributions of Modal Response Frequencies

	A	B	C	D
90–99	2	1	2	2
80–89	1	1	5	4
70–79				2
60–69	4	2	1	1
50–59	3	6	2	1
N	10	10	10	10
Mean	67	62	80	82

for the ten-item scale; (2) phi coefficient indicating correlation between item response and total score for the ten-item scale (using top 27 per cent against bottom 27 per cent); (3) the median inter-item correlation between an item and the other nine items in the scale, using tetrachoric r's; and (4) the modal response frequency of items (i.e., the percentage of respondents who answered the item in the same way). The odd-even reliability (estimated from the Spearman-Brown prophecy formula) is as follows for each of the ten-item scales: A, $+.90$; B, $+.91$; C, $+.83$; and D, $+.86$.

Results

The relative merits of each of the three methods of item selection and scale refinement are discussed below in terms of the data presented.

Scale A (Inter-Item Tetrachoric Correlations). The selection of ten items from the pool of 62 items so as to maximize the median inter-item tetrachoric correlation coefficient produced a scale which does not quite meet Guttman's criterion of 90 per cent reproducibility (see Table 1). However, this scale does compare favorably with the other scales when examination is made of the phi coefficients for each item versus the total score, and of the modal response frequencies of the ten items. Thus, the use of this method of item selection yields a relatively good scale in spite of the fact that the tetrachoric r is not an appropriate statistic to use with data of this kind. Had a more appropriate measure of correlation been used, one would assume that this method for selecting items would have yielded the best scale of the four.[1] For the writers to have used another statistic would have made the labor and expense of computation with a matrix of 62 items prohibitive. The writers therefore fell back on the same solution used by many others and resorted to the Thurstone et al., tetrachoric computing diagrams as a ready, and reasonably approximate estimate of the relationship between items. Some items, however, have extreme response splits, so that the value of r could only be estimated, or became a meaningless value of plus or minus 1.0.

Scale B (Top versus Bottom 27 Per Cent). Comparing the item responses of extreme top and bottom groups fails to work as a method of producing a scale having unidimensionality, as defined by Guttman. It does, however, produce a scale having high internal consistency as measured by the odd-even reliability coefficient (.91), or median item versus total score phi coefficient (.79). Its items, moreover, discriminate well

---

[1] A discussion of the disadvantages of the use of tetrachoric correlations with attitude scale items is discussed in Gage (7). That tetrachoric r's give different values than are obtained with other statistics was demonstrated empirically for one ten-item matrix. Greatest discrepancies occur when the modal response frequency approaches 100 per cent.

over a wide range, being more satisfactory in this respect than the items producing higher reproducibilities. This method has the additional advantage of being less laborious, and of involving less judgment and more mechanical selection of items, than the Guttman methods.

Scales C and D (Reproducibility). If one accepts Guttman's definition of unidimensionality of a scale, one requires among other things that the scale have a per cent reproducibility of 90 per cent or more. The only scales which meet this requirement are scales C and D. Furthermore, practically the same results were obtained when items were selected from a pool of 62 heterogeneous items as when from a much more homogeneous group of 27 items. These results obtain even though the Cornell Technique of Scale Analysis is not designed primarily as a method of item analysis and item selection, and even though it is not intended to be used in the mechanical fashion in which it was used in this study.

The use of scale analysis for selecting items does have some disadvantages, however, in terms of the response distributions of items. Scales C and D selected items which were answered the same way by a large proportion of the respondents (80.2 per cent and 81.6 per cent). Moreover, scales C and D are inferior to the other two scales in terms of odd-even reliability.

Discussion

It has been the purpose of the present paper to present the results of an application of the Cornell Technique of Scale Analysis to an attitude scale in order to compare its workings with those of two methods which have been heretofore considered appropriate for scale refinement. There are certain side issues which come up in the use of the Cornell Technique which complicate its use in such circumstances. The chief obstacle is that one must consider several features of an item at the same time in manipulating data for analysis. For instance, the fewer response categories an item has, the easier it will be to "reproduce" that item's response, knowing an individual's total score on the scale. A scale made up of dichotomized items thus has higher reproducibility than a scale with the same items with three or more responses. Also, the prediction of an individual's response to a particular item can be made with greater accuracy if a very high percentage of the total group answer that item in the same way. A scale made up of only very popular and very unpopular items will, therefore, have higher reproducibility than one made up of items of varying degrees of popularity. To avoid spuriously high reproducibility resulting from many items of this sort, Guttman has set up the requirement that no category have more error in it than non-error. Thus when 90 per cent agree with an item in a scale, we must

predict correctly, half of the time, from total score, not only who the 90 per cent are who say agree, but who the 10 per cent are who disagree.

It is difficult to process one's data keeping these various requirements in mind. (One wishes that the mechanics of scale analysis could receive the sort of synthesis and organization which the Wherry-Doolittle method provides in solving the problems of multiple regression.) In addition, one finds that the safeguards invented by Guttman occasionally permit worthless items to remain in the pool. Most serious weakness is in the more-error-than-non-error-per-category rule. It is possible to have an item with 99 per cent reproducibility which meets this rule, which is nonetheless worthless in that it has zero relation to the total score on the scale. If all but two persons agree with an item, and one of these has the highest total score and the other the lowest total score, then the item is valueless but meets the requirements Guttman sets forth.

Guttman's techniques cannot be used easily by research workers who have not had considerable experience with them.[2] Much judgment must be exercised in the combining of response categories and in balancing the several criteria of unidimensionality which Guttman has developed (reproducibility, more error than non-error in each category, items selected at various intervals along the range of modal response frequencies). Special care must be taken to avoid the selection of too many items with high modal response frequencies, since such items, while having high per cent reproducibilities, tend to have low reliability and low discriminating power.

Thus in one sense, Guttman's approach is a less satisfactory approach to the problems of scale refinement than the traditional methods. The worker is required to judge, first of all, whether or not items can logically be considered to belong together.[3] He must then scrutinize the pattern of responses of individuals to each of these selected items in terms of total scores on the scale and decide how best to combine item response categories in order to improve the scale. Throughout the entire analysis, there are no rigorous tests applied to determine which of several methods will work best. In fact, the worker must keep in mind several different item characteristics while he works.

In spite of the mechanical difficulties of scale analysis, however, the writers find it a valuable and useful technique. The judgmental processes mentioned above do have the beneficial effect of compelling the investigator to become better acquainted with the data with which he

[2] Edwards (5) has shown that even Guttman's own published data may be reworked to yield different results than originally reported.

[3] Edwards and Kilpatrick (6) have described a method more precise than Guttman's for selecting items which will constitute a unidimensional scale using Guttman's criteria of unidimensionality.

works. The forcing of judgments on the worker constantly takes him back to the data themselves and this is highly desirable. Moreover, there are advantages in predicting a response from total score instead of the reverse, and in predicting from the total score instead of predicting the response to one item from the response to another item. Consider a scale which obviously has perfect unidimensionality; for instance, the questions: Are you over 10 years old? Are you over 20 years old? Are you over 30 years old?, etc. Knowing the total score, the responses to every item can be reproduced with perfection. Knowing the response to only one item, one may or may not be able to predict the responses to all of the other items, and one cannot, therefore, always predict the total score without error. High reproducibility, therefore, has more meaning in defining the unidimensionality of a scale than either high item-versus-total score correlations or high item-versus-item correlations (4).

Finally, one must avoid thinking of scalability, as defined by Guttman, as a "good" characteristic of a series of items and of non-scalability as a "bad" characteristic. The use which is to be made of the series of items must always be considered. If the measure in question is to be used as a predictor variable for instance, scalability may be irrelevant or even undesirable. If the measure is to be used in a study of mental or personality organization (perhaps as a measure of what Gordon Allport has called a "common trait"), it should represent but one dimension, and, hence, should be scalable. The measurement of public opinion also makes profitable use of scales having high reproducibility.[4]

In summary, the writers feel that Guttman's new scale analysis techniques can prove to be very useful in problems of psychological measurement.[5] Considerable discretion must be exercised, however, both in the selection of suitable problems to which these methods may be applied and in the way the methods themselves are handled.

References

1. Cheshire, L., Saffir, M., and Thurstone, L. L. *Computing diagrams for the tetrachoric correlation coefficient.* University of Chicago Book Store, Chicago, 1933.
2. Clark, K. E., and Kriedt, P. H. An application of Guttman's new scaling techniques to an attitude questionnaire. *Educ. psychol. Measmt.*, 1948, 8, 215–224.

[4] For a further discussion of the importance of scale reproducibility, see Coomb's analysis of the "trait status" score (3).

[5] In the present article the writers have attempted to call attention to the advantages and disadvantages of scale analysis methods in terms of the results obtained when these methods are used. That these methods do not, in practice, live up to the promise they show in theoretical terms may be due in part to the way in which scale analysis is done. For a discussion of this point see Clark and Kriedt (2).

3. Coombs, C. H.   Some hypotheses for the analysis of qualitative variables.   *Psychol. Rev.*, 1948, 55, 167–174.

4. Dodd, S. C.   A simple test for predicting opinions from their subclasses.   *Int. J. Opin. Attitude Res.*, 1948, **2**, 1–25.

5. Edwards, A. L.   On Guttman's scale analysis.   *Educ. psychol. Measmt.*, 1948, 8, 313–318.

6. Edwards, A. L., and Kilpatrick, F. P.   A technique for the construction of attitude scales.   *J. appl. Psychol.*, 1948, **32**, 374–384.

7. Gage, N. L.   Scaling and factorial design in opinion poll analysis.   *Purdue Univ. Studies in Higher Educ.*   LXI, 1947.   pp. vi + 87.

8. Guttman, L.   The Cornell technique for scale and intensity analysis.   *Educ. psychol. Measmt.*, 1947, **7**, 247–280.

9. Jurgensen, C. E.   Table for determining phi coefficients.   *Psychometrika*, 1947, 12, 17–29.

The Airline Pilot's Job *

Thomas Gordon

American Institute for Research, Pittsburgh, Pa.

It is the purpose of this paper to report certain aspects of a study conducted by the Aviation Branch of the American Institute for Research under the auspices of the National Research Council Committee on Aviation Psychology.[1] Funds for the project were furnished by the Civil Aeronautics Administration. This study, completed in November, 1947, was undertaken (1) to study current methods of selecting and evaluating the airline pilot and (2) to determine the critical requirements of his job. It was intended that the data obtained in this investigation be used as a basis upon which to develop improved procedures for selecting, training, and certifying airline pilots. At present the American Institute for Research is utilizing the data as a basis for devising a radically new type of flight examination for pilots seeking the Airline Transport Rating certificate. This latter project is under the same sponsorship as the study to be described in this paper.

In the first phase of the study the general procedure followed was to survey the available sources of information pertaining to present methods of selecting and evaluating airline pilots. In the second phase of the project the procedure was to survey sources of information about the critical requirements of the airline pilot's job, an attempt being made to answer the question: "What behavior and characteristics are required for handling the job safely and effectively?"

Methods of Selecting the Airline Pilot

Methods of selecting applicants for the job of airline pilot were studied by examining the personnel records of 432 pilots from five major airline companies. The technique employed was to obtain the records of pilots who had been released by their companies because of lack of flying proficiency during the period between initial hiring and the time when they

* Parts of this paper were read at the Meeting of the Aero-Medical Association in Toronto, Canada, on June 17, 1948. The author's report of the entire study has been published as Research Report No. 73 by the Civil Aeronautics Administration, Division of Research, Washington, D. C. (3).

[1] The writer is indebted to the members of this committee and to John C. Flanagan for their guidance during the study and to the members of the Aviation Branch of the American Institute for Research who assisted in conducting the study.

122

would have qualified as an airline captain. These pilots constituted the experimental group (E-group). Then the records were obtained on a number of pilots who had not been eliminated but were currently employed. These pilots constituted the control group (C-group). The two groups were matched on the basis of time of original hiring by the company. Adequate data were available for both the experimental and control groups on eight variables of the type currently established by airline companies as selection requirements for pilot applicants. These variables were: (1) Age at time of hiring; (2) Previous education; (3) Otis Test I.Q. scores; (4) Bennett Test of Mechanical Comprehension (Form AA) scores; (5) Minnesota Multiphasic Personality Inventory scores; (6) Previous flying hours; (7) Marital status; and (8) Previous ground training in aeronautical subjects.

The experimental group and the control group were compared on each of these variables. Data were not available for all of the pilots in each group on each separate variable. The findings, summarized in Table 1, show that the difference between the group of eliminated pilots (E-group) and the group of successful pilots (C-group) on no one of the eight variables was statistically significant even at the 5% level of significance. These results indicate rather conclusively that present requirements established by airline companies for selection of applicants are not adequate for predicting later success or failure with much confidence. Furthermore, because none of the selection procedures differentiated between eliminated and successful pilots it was not possible to derive from these procedures any clues as to the critical requirements of the pilot's job.

Methods of Evaluating the Airline Pilot

The methods and procedures used by airlines for evaluating their pilots also were surveyed in this study. Information pertaining to evaluation procedures was obtained primarily through examination of company records of the flight performance and ground school achievement of both eliminated and currently employed pilots and through scrutiny of the flight examinations used by airlines. Pilots' and check-pilots' attitudes toward present methods of evaluation and their suggestions for improvement were obtained through individual interviews. The findings in regard to methods of evaluation currently used by airline companies can be summarized as follows:

1. There exists a great amount of variation between airline companies as to the adequacy of the training records maintained on their pilots. There were practically no records of flight tests in the files of some of the pilots.

Table 1

Comparison of Eliminated and Successful Airline Pilot Trainees on Selection
Requirements Established by Airline Companies

	Number of Pilots		Mean Difference (C minus E)	Standard Error of Difference	t-ratio*
Selection Requirements	E-group	C-group			
1. Age at Time of Employment	169	166	− .65 yrs.	.50	1.297
2. Amount of Education at Time of Employment Beyond High School	170	169	.18 yrs.	.24	.738
3. Otis I.Q.'s	63	63	2.40	1.36	1.762
4. Bennett Test of Mechanical Comprehension Scores (Form AA)	14	14	6.14	7.24	.848
5. Minnesota Multiphasic Personality Inventory Scores:					
Lie Scale	16	16	1.00	1.77	.566
Validity Scale	15	15	.33	1.13	.294
Hypochondriasis Scale	15	15	−1.50	1.49	.222
Aggression Scale	16	16	1.00	2.91	.514
Hysteria Scale	18	18	1.00	1.81	.553
Psychopath. Deviate Scale	15	15	3.47	3.07	1.130
Interest Scale	17	17	−1.71	3.03	.564
Paranoia Scale	17	17	1.29	1.65	.780
Psychasthenia Scale	15	15	.00	2.29	.000
Schizophrenia Scale	15	15	− .87	1.60	.544
Hypomania Scale	16	16	2.81	1.44	1.957
6. Number of Previous Flying Hours	165	171	14.7 hrs.	128.92	.113
7. Marital Status	170	168	(117 married in E-group, 119 in C-group)		
8. Previous Training in Aeronautical Subjects	214	214	(134 with previous training in both E-group and C-group)		

* None of the mean differences were statistically significant at the 5% level of significance.

2. All of the airlines rely upon periodic flight examinations, called flight-checks, for obtaining evaluations of their pilots. The maneuvers that make up the flight-check vary from one airline to another and vary somewhat within a single airline from one check-pilot to another.

3. In general, on these flight-checks a pilot is rated against the standard set up by the particular check-pilot rather than against an objective standard. For example, pilots are usually rated on a scale, such as "Standard-Substandard," "Good-Average-Below Average," or

"1-2-3-4-5." For a few maneuvers, however, some airlines have tried to establish more objective standards in the form of "limits" for altitude, airspeed or heading, within which the examinee is required to keep the plane in order to achieve a passing rating. Even the application of these limits was found to vary among check-pilots within a single airline.

4. The "halo effect" was found to be operating in the ratings of flight performance. From examination of the records of past flight-checks, it was found that when a pilot received a below-average rating on one maneuver there was a very strong tendency for him to get below-average ratings on all subsequent maneuvers. The ratings did not differentiate between the pilots' strengths and weaknesses. In other words, they are not useful as diagnostic performance records. One explanation of the lack of discrimination of the ratings on different maneuvers is the fact that it is common practice to make no record of the pilot's performance during the flight. The flight-test forms are usually filled out after the flight, thus increasing the chance that the quality of a pilot's performance on specific maneuvers might be forgotten.

The discovery of the inadequacy of records of performance, the lack of standardized evaluation procedures and the subjectivity of the measures of proficiency parallels the findings of Army Air Force psychologists whose research in the area of pilot proficiency measures is reported in the volume edited by Miller (5). Similar findings have been reported in a study of the flight-checks used by the Civil Aeronautics Administration (1).

The Critical Requirements of the Job

A second objective of the study was to determine the critical requirements of the job of airline pilot. This involved an analysis of the job with particular emphasis upon isolating those job requirements which are the most critical. In this approach, "critical requirements" are defined as those job requirements, expressed in behavioral terms, which have proved to be important factors in differentiating successful or unsuccessful performance on the job. The assumption underlying this approach is that the most critical differences between the safe and effective pilot and the one who is not will be revealed by focusing the job analysis upon situations where the behavior of pilots has been shown to make a difference. To use a common expression of pilots, the critical requirement approach attempts to determine "what separates the men from the boys." Flanagan (2) has described the use of this job analysis method in the study of causes of mission failures in the Army Air Forces, and he has stated that such a determination of critical requirements is the principal objective of job analysis procedures.

Although this approach resembles other methods of job analysis, it also differs from them in an important respect. It is common practice in most job analysis approaches to collect long lists of job requirements, after which it is necessary to submit them to experts (usually psychologists, supervisors, top management) for judgments of their relative importance for success on the job. The critical requirement approach, however, yields at the outset *only* the critical requirements, and it relies more upon the participants on the job for judgments of what is critical or upon actual records of situations where behavior has been critical. For example, in this study we relied upon the following sources of information:

1. An analysis was made of the records of all scheduled domestic airline accidents, during the period of 1938 through 1946, in which the behavior of the pilot was judged a contributing factor in the accident. From each of 121 such accident reports we extracted a description of the specific behavior of the pilot prior to and during the accident and the circumstances leading up to the accident.

2. Interviews were conducted with airline pilots and check-pilots for the purpose of obtaining a larger sample of critical incidents than was provided by the accident reports. Questions were devised which would yield examples of critical situations rather than commonplace or everyday occurrences. This "critical incident technique" required the pilots to recall recent events or incidents in which they did something which created an unsafe situation, thus minimizing discussions of traits or stereotyped opinions as to the requirements of the job. Examples of "critical incident" questions used are:

a. "Probably all pilots who have flown a lot have done something at one time or another that got them into an uncomfortable situation or even a near-accident. We would like to get several examples of such things you have done. First, could you describe the most recent situation in which you did something like this and tell me just what you did?"
b. "Now, I would like for you to recall the last time you had to take over the controls from a co-pilot because you felt the situation was pretty critical. Could you describe that situation and tell me just what the co-pilot did or might have done if you hadn't taken over?"
c. "We would like to draw on your experience as a check-pilot to get examples of what pilots do on check-rides. Would you think back on the last pilot you failed on a check-ride and tell me exactly what he did which caused you to fail him?"

From questions such as these we obtained 333 usable incidents from 270 interviews. Interviewing was done in 18 cities with pilots from 27 different scheduled and non-scheduled airline companies. The pilots were selected in a fairly random manner. The determining factor for selection generally was the presence of the pilots at the airport in prepara-

tion for a flight or at the completion of a flight on the particular days the interviewers visited the airport. The questions were standard for each interview and the interviewers wrote down the responses of the pilots on standard forms. An example of the kind of incidents obtained is the following:

"On daytime flight from New York to Miami in DC-3, the weather was clear with wind gusts up to 50 mph. They were landing at Raleigh with approximately 35 mph wind about 45° across runway. The co-pilot was landing the plane from the left seat. He came in too slow and was just about to touch the runway going sideways and with the downwind wing dangerously low. The captain was afraid the co-pilot might land hard enough going sideways to buckle the landing gear or get the wing low enough to cause a ground loop. The captain took the controls, added power, corrected for drift and landed OK. The captain stated that his co-pilot was inexperienced."

The next step in the analysis involved extracting from each incident the specific pilot acts contributing to the accident or near-accident. For example, in the incident above, the following critical pilot acts were extracted: (1) Executed landing approach at too low an airspeed and (2) Drifted or was not aligned with runway during round-out. The above step yielded 787 specific pilot acts, each of which had contributed to an accident or an unsafe situation. This group of acts was subjected to a further analysis in which all the acts were sorted into 21 smaller groups or clusters of homogeneous acts. These clusters made up or defined the critical components of the job of airline pilot. For example, it was found that there were four categories of errors all having to do with the operation of controls and switches: 41 instances in which pilots forgot to operate a control or switch, 31 of confusing two controls or switches, 14 of improperly adjusting a control or moving a switch in the wrong direction, and 6 of inadvertent operation of a control or switch. These four different kinds of errors of operating controls and switches formed a cluster which defined a specific job component. Twenty-one such components were extracted from the data.

As would be expected, it was found that certain components ranked higher than others as judged by the frequency of the specific pilot errors classified in the particular components.

The components of the airline pilot's job found most critical are shown in Table 2. In column (a) are listed the 21 critical requirements or components of the job which were obtained by classifying or grouping similar pilot errors extracted from three sources: (1) from analysis of airline accidents, (2) from analysis of incidents or near-accidents experienced by airline pilots, and (3) from analysis of pilot errors reported by check-pilots as reasons for failing pilots or for taking over controls from pilots on check-rides or flight examinations. The frequencies with which errors

Table 2

Critical Requirements of the Job of Airline Pilot Determined by Frequency of Errors Extracted from Accident Reports, Critical Incidents and Flight-Checks

Critical Requirements (a)	Frequency of Errors			
	(b) Acci- dents	(c) Inci- dents	(d) Flight- checks	(e) Total
1. Establishing and maintaining angle of glide, rate of descent, and gliding speed on approach to landing	47	41	11	99
2. Operating controls and switches	15	44	33	92
3. Navigating and orienting	4	39	19	62
4. Maintaining safe airspeed and attitude, recovering from stalls and spins	11	28	18	57
5. Following instrument flight procedures and observing instrument flight regulations	5	27	13	45
6. Carrying out cockpit procedures and routines	7	31	4	42
7. Establishing and maintaining alignment with runway on approach or takeoff climb	3	31	5	39
8. Attending, remaining alert, maintaining lookout	14	23	1	38
9. Utilizing and applying essential pilot information	0	19	18	37
10. Reading, checking and observing instruments, dials and gauges	1	26	7	34
11. Preparing and planning of flight	2	27	3	32
12. Judging type of landing or recovering from missed or poor landing	1	23	8	32
13. Breaking angle of glide on landing	1	25	5	31
14. Obtaining and utilizing instructions and information from control personnel	3	21	0	24
15. Reacting in an organized manner to unusual or emergency situations	0	17	7	24
16. Operating plane safely on ground	7	15	1	23
17. Flying with precision and accuracy	0	7	15	22
18. Operating and attending to radio	0	7	10	17
19. Handling of controls smoothly and with coordination	0	6	8	14
20. Preventing plane from undue stress	0	5	7	12
21. Taking safety precautions	2	5	4	11

were obtained from each of these three sources are shown in columns (b), (c), and (d). The total frequencies of errors from all sources are shown in column (e).

Table 3 presents the correlations between the rank order of the critical requirements as determined from the frequencies of pilot errors obtained

from the three sources. These were computed in order to answer the question: "To what extent do you obtain similar indices of the relative 'critical-ness' of the various job requirements from analyses of critical behavior in accidents, in incidents or near-accidents and on flight-checks."

Table 3

Correlations Between Rank Order of Critical Requirements as Determined from Three Sources of Pilot Behavior
(Spearman Rho Coefficients)

	Incidents	Flight-checks
Accidents	.71 (S.E. = .12)*	—.04 (S.E. = .23)
Incidents		.28 (S.E. = .22)

* An r of .43 is necessary for significance at the 5% level and an r of .55 for significance at the 1% level (4).

The high positive correlation between the rank order of the critical requirements determined from analysis of airline accidents and from analysis of the incidents reported by pilots indicates that with the critical incident technique we accomplished the objective of obtaining job requirements which are critical from the standpoint of safe flying. The low negative correlation between the rank order of the critical requirements as determined by analysis of accidents and as determined by analysis of behavior of pilots on flight-checks might be interpreted as an indication that check-pilots' reasons for failing pilots on flight-checks are not closely related to the requirements of the job which seem to make a difference between safe and unsafe airline flying. It may well be that the present flight-checks do not provide an adequate evaluation of the extent to which pilots demonstrate proficiency in the *most critical* aspects of airline flying. Check-pilots seem to be emphasizing proficiency in different aspects of the job, such as flying the plane smoothly and keeping within very precise limits of altitude, airspeed and heading. These requirements, of course, are probably important from the standpoint of the comfort of passengers, but at the same time they shouldn't be emphasized at the expense of neglecting requirements which are more critical from the standpoint of safety.

Summary

The lack of statistically reliable differences between eliminated and successful pilots indicates that present methods of selection do not predict success or failure in training. To achieve this, it is probably necessary for airlines to utilize new procedures which have been validated against

this or some similar criterion, rather than rely on standardized tests and interview procedures which, although of possible usefulness for predicting success in other fields, have not been validated as predictors of success in airline piloting. Furthermore, it would appear from this survey that training and proficiency records of airline companies are inadequate for use as criteria of proficiency or for providing a diagnostic picture of the proficiencies of their pilots. The findings also suggest that the subjective type of flight-check currently employed by airlines cannot adequately provide an objective evaluation of the extent to which pilots meet the most critical requirements of the job.

From the results of the analysis of pilot errors extracted from various critical incidents it has been shown which components of the pilot's job are most critical from the standpoint of safety and effectiveness on the job. The implications of these results are rather obvious, but some may be mentioned briefly:

1. The most critical requirements should receive more emphasis by check-pilots who evaluate the proficiency of pilots on flight examinations. This study suggests that special emphasis is needed on the landing approach, accurate operation of controls, methods of navigating and orienting, maintaining safe airspeeds, compensating for drift. These data are now being used as a basis for developing a more objective flight examination, as mentioned earlier.

2. The findings suggest a need for improved cockpit design to simplify the pilot's job and reduce the possibility of such errors as: confusing two contro... .aking improper adjustments of controls and inadvertently operating controls.

3. The findings might be used to suggest which components of the job need greater emphasis in the training program for pilots. Pilot trainees also could be informed which errors are most frequently made in airline flying.

4. The list of critical requirements should prove useful in devising improved methods of selecting pilots, inasmuch as they provide valuable clues as to the critical aptitudes needed by a safe airline pilot.

Finally, it would seem that the results of this study furnish evidence that the "critical incident technique" is a very useful method of isolating the critical requirements of a particular job. This method increased the size of the sample of critical incidents, which if restricted to accidents alone would have been too small to yield a sufficient number of pilot errors upon which to base the list of critical requirements.

References

1. Festinger, L., Kogan, L. S., Odbert, H. S., and Wapner, S. *An analysis of inspectors' ratings of check-flights as recorded on ACA 342Z.* Washington: CAA Division of Research, Report No. 58, March 1946.
2. Flanagan, John C. (Ed.). *The aviation psychology program in the Army Air Forces.* Washington: U. S. Government Printing Office, 1948. (AAF Aviation Psychology Program Research Report No. 1.)
3. Gordon, T. *The airline pilot: a survey of the critical requirements of his job and of pilot evaluation and selection procedures.* Washington: CAA Division of Research, Report No. 73, November 1947.
4. Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1942.
5. Miller, N. E. (Ed.). *Psychological research on pilot training.* Washington: U. S. Government Printing Office, 1947. (AAF Aviation Psychology Program Research Report No. 8.)

Factors Related to Life Insurance Selling *

D. F. Kahn and J. M. Hadley

Division of Applied Psychology, Purdue University

The purpose of the study has been three-fold: first, to determine the degree of relationship that exists between relative success in the early period of selling life insurance and success at a later period; second, to examine various selling activities with a view to uncovering certain factors which differentiate successful from unsuccessful agents, and to select such factors as might contribute to the refinement of life-insurance training programs; third, to investigate further certain personal history items and personality traits already known to correlate with success in selling life insurance, and to analyze other measurable areas of personality, with the aim of increasing the sensitivity of existing selection methods. The identification of individuals for whom the likelihood of success is known would not only benefit management, but would, to some extent, minimize feelings of frustration on the part of the agent who, from the outset, may be doomed to failure.

Procedure

The subjects considered in the present investigation were a group of 84 new life insurance agents who had attended Class I and Class II of the Purdue Course in Life Insurance Marketing (1). Each subject selected for study had received a five-week basic course at the school and had also completed thirteen weeks of selling in the field. Production records for the most part were collected during the year 1946.[1] The salesmen represented 19 life insurance companies and 22 states (3). Well over

one-half of the group were veterans attending the school under public laws affording veterans educational advantages.

Data were collected in three major areas: selling activities, personal history items, and psychological measures.

Selling activities included: (1) size of application written, (2) number of calls made. A call is defined as a face-to-face conversation with a potential buyer. (3) Number asked to buy. "Asked to buy" is defined as a salesman's discussing life insurance with a client where the latter is asked to take action toward securing a policy, and (4) number of applications written. An application is defined as a signed application for a life insurance policy.

Relationships between calls, "asked to buy," and applications written were investigated under the following headings: (1) percentage of applications written to total number of calls made, (2) percentage of persons asked to buy to total number of calls made, (3) percentage of applications written to total number of persons "asked to buy." Because of the unequal lengths of reported field work for Class I and Class II, 37 and 39 weeks respectively, and further because some of the agents, for one reason or another, withdrew from the course before completion, the measures, calls, "asked to buy," number of applications written, and production were computed on the basis of a weekly average.

During the initial five-week training period a personal history questionnaire and a battery of psychological tests were administered. Personal history items analyzed were: (1) age, (2) number of dependents, (3) living expense per month, and (4) life insurance owned, including National Service Life Insurance.

Psychological measures investigated included: (1) Kuder Preference Record (5), (2) Guilford-Martin Personnel Inventory Number I (2), (3) the previously mentioned test, used to measure degree of uncertainty as determined from the number of questions responded to as undecided, "?", (4) The Adaptability Test (8), (5) Part II[2] of the Aptitude Index (7).

The measure adopted in the present study as the criterion of success was the production records on file with the Purdue insurance school; these records were based on the total of signed *applications*, that is, written business, and not the total of business signed, examined and paid

[2] Letter grades reported on the Aptitude Index refer to Part II only, and should not be confused with the letter grades usually cited for the entire instrument, and which are derived from an age-weighted combination of Part I and Part II of the Index. A special form, Part III, that is to say, the Personal History portion of the Aptitude Index which was devised for use with former service men, was administered but not incorporated into this study because of the inaccuracy and the incompleteness of response to it. Three of the items appearing on Part I of the Index, number of dependents, living expenses per month, and amount of life insurance owned, however, have been considered separately in this study.

for, which is generally referred to as paid-for business. Varying lengths of school attendance necessitated reducing total production to average weekly production in order to evaluate the relative success of each agent.

In an endeavor to ascertain the relationships between early production and later production, two correlations were computed. The first of these considered the relationship between the average weekly production for the first 13 weeks and the average weekly production for the time spent in the field over and above those 13 weeks. Two agents who failed to report to the school after the thirteenth week were eliminated from this correlation, and thus 82 agents were left who had completed from 15 weeks to the maximum school period of 39 weeks in the field. Approximately 69 per cent of this group had completed the course. The second correlation computed was a measure of the same type of relationship; however, this time only those salesmen who had reported a minimum of 26 weeks were considered. It was felt that whatever relationship would be found to exist in the latter correlation would be a more accurate reflection of differences between early and late selling, since the first 13 weeks would be compared with a selling period of equal or greater length. Sixty-five cases meeting such a requirement were found, approximately 87 per cent of whom had completed the course.

In order to fulfill the second and third purposes of this study, that is, to determine whether or not the measures selected would differentiate successful from unsuccessful salesmen, two such contrasting groups of agents were identified. In making such a distinction, the agents were, in the first place, ranked from high to low on their total sales while attending the school. The production records covering the duration of the school term for those agents who, for one reason or another, did not complete the course, but did continue to sell insurance, were obtained from the respective agency managers under whom the subjects were selling. These records were used as checks against the agents' weekly reports of production made to the school, and were thus used to substantiate the original rank order of the agents in the study. Six of the group of 84 salesmen withdrew from the school at an early date after the first 13-week period, and hence their records were incomplete. These agents had either terminated with their companies, or had continued as life insurance salesmen, but for various reasons further records on their selling activities were not made available. It was felt that the inclusion of these men in the analysis would, to some extent, invalidate the rather stable ranking of the remaining 78 agents. For these reasons six agents were omitted from this part of the study. The ranked average weekly production records were divided into three equal groups, numbering 26 agents each. The high and the low groups (average weekly production $8,602 and $2,181 respectively) were designed as successful and unsuccessful.

In an attempt to locate new test items which might be of value in future selection devices an item analysie was undertaken on the 150 questions appearing on the Guilford-Martin Personnel Inventory Number I. The D-value method based on Lawshe's nomograph (6) adapted from the Kelley technique (4) was employed in this part of the procedure. All items responded to by the agents as uncertain, "?", were grouped with the "No" responses.

Results

A correlation of $+.61$, based on the records of 82 agents, was obtained between average weekly production for the first 13 weeks and average weekly production for a period of from two to 26 weeks beyond the initial period. For the group of 65 agents who had completed at least a second 13-week selling period, approximately 87 per cent of whom had reported records for the entire 37 or 39 week course, a correlation between the average weekly production for the first 13 weeks and the average weekly production for at least a second 13 weeks was found to be $+.55$. Both of the above-mentioned correlations are significant beyond the one per cent level of confidence.

Analysis of the selling activities measured revealed that several significant differences existed between the groups of successful and unsuccessful salesmen. The successful salesmen were higher in every comparison except the percentage of prospects asked to buy. They asked more people to buy but this can be explained by the fact that they averaged more calls per week. Although both groups of agents asked approximately 37 persons to buy insurance out of each 100 calls made, the successful salesmen sold insurance to approximately 31 per cent of such prospects as contrasted with the unsuccessful salesmen who sold to approximately 17 per cent.

Although success in insurance selling is, in part, determined by the denomination of the applications written by a salesman, other factors are also important. While it is true that the average size of policy written by the high and low groups of this study is different, the difference between these averages only partially accounts for the difference between the two groups with respect to the average weekly production figures. Analysis revealed that the successful salesmen were actually able to sell to a significantly larger percentage of persons called upon. In view of the average number of applications written per week by both groups, the successful group of salesmen would have been able to produce over twice as much insurance written as the unsuccessful group, even if the size of the application written had been exactly the same for both groups. The successful group was therefore able, in terms of the number of persons to whom they sold alone, to do more selling than the unsuccessful group.

Table 1

Comparisons of Mean Differences Between High and Low Producing Groups of Agents in Various Selling Activities and Personal History Items

Item	Mean of High Group	No. of Cases High Group	Mean of Low Group	No. of Cases Low Group	Mean Diff.	S.E. of Diff.	C.R.
Average Weekly Production (in dollars)	8,602	26	2,181	26	6,421	546	11.76
Size of Application (in dollars)	5,958	26	3,538	26	2,419	784	3.08
Average No. of Applications per Week	1.66	26	.74	26	.93	.14	6.70
Average No. of Cases per Week	18.68	26	13.77	26	4.91	1.65	2.98
Average No. Asked-to-Buy per Week	6.86	26	5.23	26	1.63	1.02	1.60
% of Applications to Total No. of Cases	9.66	26	5.81	26	3.84	1.00	3.84
% Asked-to-Buy to Total No. of Cases	37.50	26	37.82	26	−.31	4.70	.07
% of Applications to Total No. Asked-to-Buy	30.87	26	17.16	26	13.71	3.87	3.54
Age*	30.27	26	28.35	26	1.92	1.70	1.13
Dependents*	1.46	26	1.15	26	.31	.30	1.02
Monthly Living Expenses (in dollars)*	209	26	187	23	22	19	1.11
Life Insurance Owned (in dollars)*	16,544	25	9,158	24	7,386	1,587	4.65

* At entry into life insurance business.

Differences between the two groups in question with respect to personal history items revealed, as may be seen in Table 1, that of the four items analyzed only one, namely the amount of life insurance owned at the time of entry into selling, resulted in a critical ratio significant beyond the one per cent level of confidence. Nevertheless, the average agent in the successful group was found to be older, to have a greater number of dependents, and to have a higher standard of living as determined by living expenses per month. It is believed that investigation into personal history items that appear among various selection devices would reveal that the age factor is closely related to several other items commonly employed in typical questionnaires such as number of dependents, amount of insurance owned and so forth.

Table 2 shows differences between mean scores for the high and the

low producing groups of agents for each of the nine areas dealt with by the Kuder Preference test, and gives further the critical ratios for the significance of the differences between the means obtained in these areas. The highest critical ratio, 2.81, was obtained in the area entitled "Clerical." The average successful salesmen scored significantly lower in this area than did the unsuccessful. The critical ratio for the persuasive component was found to be only 1.92, thus significant at approximately the five per cent level of confidence.

As evidenced by the critical ratios appearing in Table 2 no differences significant beyond the ten per cent level of confidence were found to exist between the mean scores for the low-producing salesmen for traits measured by the Guilford-Martin Personnel Inventory Number I.

A critical ratio of 1.57, as shown in Table 2, resulted from a testing of the significance of the difference between the average number of question-mark responses appearing between the high and low-producing groups.

Table 2

Comparison of Mean Differences Between High and Low Producing Groups of Agents in Various Psychological Measures

Psychological Measures (Kuder Preference Record)	Mean Raw Score of High Group	No. of Cases High Group	Mean Raw Score of Low Group	No. of Cases Low Group	Mean Diff.	S.E. of Diff.	C.R.
Mechanical	63.73	26	57.87	23	5.86	5.42	1.08
Computational	29.88	26	32.87	23	−2.99	2.84	1.05
Scientific	47.38	26	45.52	23	1.86	3.57	.52
Persuasive	112.12	26	104.22	23	7.90	4.12	1.92
Artistic	41.46	26	37.17	23	4.29	3.60	1.19
Literary	48.31	26	53.70	23	−5.39	3.74	1.44
Musical	21.96	26	25.65	23	−3.70	2.30	1.61
Social Service	78.04	26	76.43	23	1.60	4.29	.37
Clerical	47.96	26	56.09	23	−8.13	2.89	2.81
(Guilford-Martin I)							
Objectivity	49.79	24	54.67	24	−4.88	2.98	1.64
Agreeableness	30.63	24	31.96	24	−1.33	2.45	.54
Cooperativeness	68.17	24	68.00	24	.17	3.92	.04
Degree of Uncertainty*	8.42	24	12.83	24	−4.42	2.81	1.57
Intelligence (Adaptability Test)	21.32	25	22.08	25	− .76	1.60	.47
Aptitude Index Part Two	46.08	26	42.28	25	3.80	2.34	1.62

* As determined by the "?" count on the Guilford-Martin Personnel Inventory No. I.

It was noted that two men in the low-producing group answered at least forty out of a total of 150 questions in the Inventory as undecided, which is an unusually large number in terms of the general distribution on the Inventory. However, what may prove to be an important finding is the fact that, when the entire group of agents is considered, the three men who received a score of 40 or over in question-mark responses, produced an average mean weekly production figure of $2,689 as contrasted with the similar average of $5,281 for the 66 agents who scored at 31 and below. There were, moreover, no scores between 30 and 40 for the whole group of agents in question. Although the number of agents in this study who scored unusually high on the measure designated as undecided is too small to allow one to place much confidence in it as an absolute finding, it is believed that further investigation along these lines with larger samples might prove fruitful.

Even though no significant differences were found to exist between the high and the low groups of salesmen when measured by the scores derived from the Guilford-Martin Personnel Inventory, an item analysis was undertaken in the hope of uncovering certain items that might possibly discriminate between the two groups of salesmen. The eight items producing the highest D-Values were: 42, 48, 77, 83, 99, 103, 135, 139. To all of these items, with the exception of item 83, the successful group of salesmen responded with a higher percentage of "Yes" answers than did the unsuccessful group of salesmen. Only four of these items were found to be significant beyond the five per cent level of confidence. These were items 42, 77, 135 and 139, having respective critical ratios of 2.73, 2.14, 2.11, and 2.82, reflecting the significance of the differences between the percentages of "Yes" responses to each item for the high and low producing groups of salesmen. However, four such items, significant at the five per cent level, would be expected to occur in a test of 150 items by chance alone. Still it is quite likely that one or more of these items might well continue to be discriminating and reliable items. Further investigation would probably shed more light on this question.

No appreciable difference, as is evidenced by Table 2, was found to exist between the average mental ability of the successful and unsuccessful groups of salesmen. A slight relationship exists between the combination of personality characteristics measured by Part II of the Aptitude Index and life insurance production.

Summary

Based solely on the criterion of written business, and pertaining only to those particular life insurance salesmen investigated in this study, the following conclusions may be drawn.

1. The degree of success during approximately the first three months offers a significantly better than chance basis for predicting the degree of success in the life insurance selling at a later date. The correlation between sales during the first 13 weeks of selling and a second period of 13 or more weeks is $+.55$.

2. Significant differences in favor of the successful agents were found to exist between the two criterion groups with respect to the following aspects:

1. Average number of calls per week.
2. Number of applications written per 100 persons "Asked to buy."
3. Number of applications written per 100 persons called upon.
4. Average size of application.
5. Average number of applications written per week.

3. Non-significant differences in favor of the successful agents were found to exist between the two criterion groups with respect to the number of persons "asked to buy" insurance per week. Since the number of persons called upon was significantly higher for the successful groups the percentage of persons "asked to buy" per 100 called upon was almost identical for the two groups of salesmen.

4. Of the four personal history items investigated, only one, namely, amount of insurance owned at entry, was found to differentiate significantly beyond the one per cent confidence level between successful and unsuccessful life insurance salesmen. The other three items, age at entry, number of dependents, and minimum living expenses per month, showed positive relationships to the criterion although no significant difference between the two groups in question was found to exist for these measures.

5. The findings of the present study indicate that the Kuder Preference Record, as commonly used, may identify life insurance salesmen but does not differentiate successful from unsuccessful agents. However, the analysis of the present data indicates that there are inherent in the Record certain relationships with success in selling life insurance that may prove to be useful in selecting high producing salesmen.

6. No significant differences between the two criterion groups were obtained for any of the three component measures of The Guilford-Martin Personnel Inventory. A supplementary measure, degree of uncertainty, as determined from the number of question-mark responses, similarly showed no significant difference to exist. One unusual finding, however, deserves mention: the three men in groups whose degree of certainty score was abnormally high were identified as producing very far below the mean of the total group. While this number is too small to

permit generalization, it is suggested that such a score may well warrant further investigation.

7. An item analysis of the 150 items of the Guilford-Martin Inventory revealed only four items which distinguished between the criterion groups significantly beyond the five per cent level of confidence. The result reflected by these four items may be considered to be well within chance expectation for a test of the present length. Nevertheless, further investigation may possibly prove one or more of these items to be serviceable enough to warrant their inclusion in a selective device. Although not a finding of the present study, it is believed possible that existing personality tests when carefully analyzed may reveal behavior patterns common to successful life insurance agents. It is also believed that unstructured or projective tests may prove of value by tapping those personality characteristics not capable of being indentified by the usual structured test.

8. No significant difference was found to exist between the mental ability test scores of the successful and the unsuccessful salesmen as measured by this tool; the mean scores of both criterion groups was for all practical purposes the same on The Adaptability Test.

9. Although no significant difference was obtained between the mean raw scores of the two groups in question, trends present in the data indicate that Part II of the Aptitude Index may have some predictive value.

Received August 19, 1948.

References

1. Barnes, D. F. *The Purdue course in life insurance marketing.* New York: The National Association of Life Underwriters, 1946, pp. 24.
2. Guilford, J. P., and Martin, H. G. *Guilford-Martin personnel inventory, manual of directions and norms.* Beverly Hills, Calif.: Sheridan Supply Co., 1943, pp. 2.
3. Kahn, D. F. *An analysis of life insurance salesmen.* Unpublished master's thesis, Purdue University Libraries, West Lafayette, Indiana, 1946.
4. Kelley, T. L. Selection of upper and lower groups for the validation of test items. *J. appl. Psychol.,* 1939, 30, 17–24.
5. Kuder, F. G. *Intermediate manual for the Kuder preference record.* Chicago: Science Research Associates, 1944, pp. 16.
6. Lawshe, C. H., Jr. A nomograph for estimating the validity of test items. *J. appl. Psychol.,* 1942, 26, 846–849.
7. Life Insurance Agency Management Association. *The value and use of the Aptitude Index.* Hartford, Conn.: Life Insurance Agency Management Assoc., 1946, pp. 24.
8. Tiffin, J., and Lawshe, C. H. *Preliminary manual for the adaptability test.* Chicago: Science Research Associates, 1943, pp. 9.

A Window-Stencil Method for Scoring the Strong Vocational Interest Blank (Men)

J. E. Greene, R. T. Osborne, and Wilma B. Sanders

The University of Georgia

The Strong Vocational Interest Blank (Men) is generally recognized by clinical psychologists and guidance workers as being one of the most useful instruments for determining the vocational interests of male counselees. The nature of the standardization of the Strong Blank is such, in our belief, as to give it a higher degree of specific validity for many counselees than that which may be obtained from other tests of vocational interest. On the other hand, many circumstances conspire to prevent as frequent and effective use of the Strong Blank as its basic validity would seem to warrant. In many counseling situations, the counselor may wish to secure immediately Strong scores on selected occupations for one or a relatively few clients. Under these circumstances local machine scoring of the test is inadvisable. Moreover, if the counselor must send the answer sheet to some off-campus test scoring service, there will be an unwanted and often crucial delay in obtaining the test results. When for either of these reasons machine scoring becomes inadvisable, the counselor must at present resort to the use of the intricate, time-consuming and error-ridden process of hand scoring the test by means of the Strong ladder stencils, or of choosing some quickly-scorable alternate test of vocational interest which often is less valid for his particular purpose than the Strong test would be.

The background for the development of the simplified scheme of hand scoring the Strong Blank herein presented may be briefly stated. In 1945 the senior author, while serving temporarily as Director of the Veterans Guidance Center of the University of Georgia, became impressed with the local need for a simplified procedure for hand scoring the Strong Blank. A large proportion of our case load consisted of male veterans interested in some type of college training. For many of these clients it was obvious that the Strong Vocational Interest Blank would provide more valid and useful measures of vocational interest than would any other instrument commercially available. Consequently, the senior author set for himself the task of devising an accurate and quick

procedure for hand scoring the Strong Blank.[1]   The basic procedure
consisted of the development of four window stencils to which were
transferred the positive and negative weights assigned to each of the
400 items of the Blank, for each Strong occupational category separately.

Since its introduction, this window stencil scoring system has been
used locally on more than 5000 cases.   Our data indicate that a semi-
skilled psychometrist can score the Strong Blank at the rate of 2½
minutes per occupation.   In our own set-up, as well as in many similar
counseling situations, the counselor typically will not need to have the
Strong scored on all possible keys.   Our experience indicates that for a
particular client we seldom wish scores on more than six of the occupa-
tional categories.   Consequently, the total amount of scoring time for
the typical client seldom exceeds fifteen minutes.   Where large numbers
of papers are to be scored on all possible keys, one of the Standard IBM
methods or the Hankes system is more economical.   In addition to
offering less opportunity for addition errors, and other errors due to faulty
alignment of the scoring stencils, the procedure herein described has the
advantage of being more economical of time and money.   For example,
the Strong ladder scoring system presupposes that a Strong booklet (8¢
each) will be expended for each client, whereas under our system IBM
answer sheets (IBM Form ITS 1100 B 360 Rev—@ 2.35¢ each) are used
and the booklet is not expended.   In terms of clerical time involved, our
window scoring stencil requires only approximately one-fourth as much
time per occupation as does the Strong ladder scoring system.

---

[1] Our earliest scheme for using window stencils was devised by the senior author.
The junior authors have subsequently refined and further simplified our earliest pro-
cedures.   For example, our original procedure required 16 separate window stencils for
each of the Strong keys, as follows:

   (a) 4 stencils for the *positive* weights on page *1* of the IBM Answer Sheet, a separate
       stencil for weight +1, +2, +3, and +4.
   (b) 4 stencils for the *positive* weights on page *2* of the IBM Answer Sheet, a separate
       stencil for weight +1, +2, +3, and +4.
   (c) 4 stencils for the *negative* weights on page *1* of the IBM Answer Sheet, a separate
       stencil for weight −1, −2, −3, and −4.
   (d) 4 stencils for the *negative* weights on page *2* of the IBM Answer Sheet, a separate
       stencil for weight −1, −2, −3, and −4.

As contrasted with our earlier procedure which employed the 16 separate window
stencils indicated above, the present system employs only 4 window stencils.   The
four separate stencils indicated under (a), (b), (c), and (d) above have each been con-
solidated into a single stencil.   As is indicated in Figure 1, all of the positive weights
for page 1 of the answer sheet are shown on the same window stencil.   Weights of +2,
+3, and +4 are indicated to the right of the respective windows; all the remaining
windows for this stencil have a weight of +1, but our experience indicates that it is
preferable not to show the weight of +1 to the right of the window.

Development of the Window Stencil System

It is proposed to describe our procedure in some detail so that persons who wish to do so may prepare their own window stencils for as many or as few of the Strong keys as may be desired in a local counseling situation. As was implied above, our basic procedure consisted of transferring from the Strong *ladder* stencils for a given occupational category (e.g., Chemist) to our own *window* stencils for that same category the several positive ($+1$, $+2$, $+3$, and $+4$) weights and negative (-1, -2, -3, and -4) weights assigned to any given response to each of the 400 items in the Strong Blank. This process of transferring weights, involving the several steps indicated below, will be illustrated with the scale for Chemist.

Step 1. Making use of page 1 (items 1–200) of the IBM Answer Sheet for Strong's Vocational Interest Blank for Men (Revised) Form M[2] for recording the value of the various weights involved, we transferred from the Strong ladder stencil for Chemist all the $+1$, $+2$, $+3$, and $+4$ weights which Strong assigned to these 200 items on the Chemist scale.[3] In the same manner, all the *positive* weights assigned to items 201–400 were recorded on page 2 of a *second* Strong Answer Sheet. Thus these two separate recordings carried all the *positive* weights assigned to Chemist in the 400 items of the Strong Blank. Similarly, all the -1, -2, -3, and -4 weights assigned to items 1–200 were recorded on page *1* of a *third* answer sheet and the *negative* weights assigned to items 201–400 were recorded on page *2* of a *fourth* answer sheet. This procedure resulted, therefore, in transferring from *9* ladder stencils (each having 3 separate columns of weights of varying size and sign) to *4* answer sheets all the positive and negative weights assigned to Chemist on the 400 items comprising the Strong test.

Step 2. The final step involved in preparing the *4* window stencils to replace the *9* ladder stencils for the Chemist scale required little time and material. In punching all of our window stencils we used the standard heavy cardboard form, International Test Scoring Machine

---

[2] IBM Form ITS 1100 B 360 Rev. Copyrighted by the Board of Trustees of Leland Stanford Junior University.

[3] In practice, this procedure of transferring weights will be facilitated if two persons cooperate in the following manner: One person will apply the ladder stencil for Chemist to the Strong Interest Blank booklet and read off the $+1$, $+2$, $+3$, and $+4$ weights assigned to each response to items 1–200. For example, "Item 1—no weight; Item 2—D, $+2$; Item 3—I, $+1$; Item 4—no weight; Item 5—no weight; Item 6—L, $+2$; Item 7—D, $+1$; Item 8—D, $+2$; Item 9—no weight; Item 10—L, $+4$; etc." The second person will record these positive weights in the appropriate spaces on page 1 of the answer sheet. The appropriate weights for Stencils B, C, and D will be determined similarly.

Key Form A.[4]   Each of the 4 answer sheets described above was used as a basis for punching a window stencil to which the weights recorded on the answer sheets were accurately assigned.   For example, the answer sheet which recorded the $+1$, $+2$, $+3$, and $+4$ values on items 1–200 was fitted *exactly* against the back[5] of one of the cardboard forms 1000 A 310 and appropriate response positions were punched through both the answer sheet and the cardboard form with a pin.   Then, working from the front (i.e., printed) side of Form 1000 A 310, each circle[6] through which the pin had been punched was converted into a "window" with an IBM hand punch.   The weight of each response to each item was indicated according to the procedure described in footnote number 1 and illustrated in Figure 1.
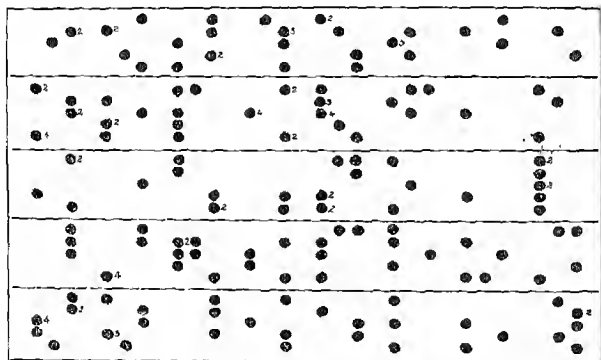
Fig. 1.   Man chemist—stencil A: plus weights, items 1–200.
(For page 1 of Answer Sheet.)

Note: If sufficient demand should develop for these window stencils arrangements will probably be made with Stanford University Press to produce them.

Thus for each occupational scale, 4 window stencils were prepared, as follows:

[4] IBM Form ITS 1000 A 310.   These forms may be procured from the International Business Machines Corporation.   Cost, 2.3¢ each.

[5] The back rather than the front of the cardboard form was used in order to reduce the amount of eye strain in scoring.   The circles on the front of the cardboard form tend to produce mental confusion and fatigue of the eye muscles.

[6] These circles were used as guides in securing accuracy in punching.

Stencil A.   Positive weights, page 1 of Answer Sheet (items 1–200);
Stencil B.   Positive weights, page 2 of Answer Sheet (items 201–400);
Stencil C.   Negative weights, page 1 of Answer Sheet (items 1–200);
Stencil D.   Negative weights, page 2 of Answer Sheet (items 201–400).

Procedures for Window-Stencil Scoring

Once the window stencils have been prepared, hand scoring of the Strong (Men) becomes greatly simplified. Obviously, to evaluate the client's interest in a given Strong occupational category, it is necessary to secure the algebraic sum of the positive and negative weights which he earned on that scale. For any given scale, the sum of his *positive* weights may be quickly determined by applying window stencil A to page *1* of the answer sheet and window stencil B to page *2* of the answer sheet. Similarly, the sum of his *negative* weights may be obtained by appropriate application of window stencils C and D. The algebraic total of these two sums constitutes his total raw score on the given occupational category. The raw score thus obtained corresponds exactly to the raw score obtainable by ladder stencil or machine scoring procedures and may be interpreted accordingly.

Evaluation

Although a considerable amount of exacting work was involved in our preparation of the window stencils herein described, it has been our experience that this labor expenditure was of minor significance in comparison to the vast and varied benefits which we have derived from their use. In budgetary terms, two types of savings have been notable: (1) a marked decrease in clerical time involved in window stencil scoring as contrasted with ladder stencil scoring; (2) use of IBM Answer Sheets instead of expendable Strong booklets has markedly reduced the per capita cost of testing materials and has thus permitted much more extensive use of the Strong test than otherwise would have been feasible. Finally, our experience indicates that the margin of error in scoring the Strong by our window stencil procedure is markedly less than that obtained when the Strong ladder stencil system is used.

A Short Test of Mental Ability

Jay L. Otis and David J. Chesler

Personnel Research Institute, Western Reserve University

A survey of 26 paper-and-pencil tests of mental ability suitable for use at the adult level, practically all of which are listed in the *Nineteen Forty Mental Measurements Yearbook* (1), showed that the range of "examination time" varied from 12 to 153 minutes. Five of these tests required 16 minutes or less. The median examination time was 32 minutes. It would seem that there are few short tests of mental ability suitable for adults—"short" in this connection being defined as approximately 15 minutes or less.

While, in general, no claims of superiority with respect to reliability or validity can be made for the short test as compared with a longer test, nevertheless the short test has demonstrated its usefulness and practicability, and, in many instances, certain advantages over the longer test. In the industrial employment office, where time is often at a premium, the short test of mental ability can yield results of more than acceptable validity with respect to the types of jobs and individuals involved. In those situations where the standards are more precise, the short test may be used to screen out those individuals who are obviously below or above the desired mental standards, so that a longer test of mental ability and tests for other functions will be reserved for those who fall within the accepted range. This is an economical procedure, both to the applicant and to the organization. The applicant who cannot possibly qualify is prevented from embarking on a lengthy testing program, and the organization is saved the time and expense involved in administering and scoring a complete test battery.

In the vocational guidance situation, the short test of mental ability has very useful application, also, in that it may be an excellent indicator of the type (e.g., "elementary," "intermediate," or "advanced") of longer test that should be administered. It is not an uncommon experience with psychometrists and vocational counselors to realize that the counselee has taken, or is in the process of taking, a test of mental ability which is inappropriate to his level. A short test of mental ability used as a "pre-test" will prevent this from happening. A knowledge of the testee's intelligence, obtained before the test battery is decided upon, is also extremely helpful in determining what special tests of aptitude and

achievement should be administered. For example, in the case of a counselee who wants to go to college, but whose pre-test shows him to be significantly below average in intelligence, tests of aptitude should be administered which are more applicable to a lower level of employment or training, and tests applicable at the college level should be omitted. The short pre-test serves other purposes in the counseling situation. If administered before the initial interview, it provides clues as to the degree to which the counselee will understand the verbal give-and-take of the initial interview. It can also be a fast and reliable determiner of the necessity for an individual rather than a group test of mental ability.

For these reasons the Personnel Research Institute of Western Reserve University initiated in 1942 a research project with the purpose of developing a short test of mental ability. The Personnel Research Institute was in an excellent position to undertake this project since it was set up to carry on personnel research in such areas as the development of procedures for employment and training of workers, as well as the development of techniques in the field of vocational guidance (2). The activities of the Personnel Research Institute solved in large part the problem of obtaining suitable populations for the standardization and validation of a new test. The result of this research is the *Classification Test for Industrial and Office Personnel* (3).

Description of the Test

The *Classification Test for Industrial and Office Personnel* is primarily a measure of mental ability at the adult level, although evidence has accumulated that it is also satisfactory for use at the high school level. It is a self-administering group test and intended for individuals who know how to read. An attempt was made to include items of approximately uniform difficulty throughout the test and to keep the difficulty level relatively low. Most group tests of intelligence present items in order of increasing difficulty. This is often discouraging to the ordinary shop or office worker. In addition, an increasing order of difficulty tends to reduce the total number of items required. In a short test of mental ability constructed on this basis many subjects reach their difficulty ceiling in a very short time (perhaps 7 or 8 minutes) so that the *effective* number of items is reduced still further. In the standardization of such a test it is usually found that the successful completion of even two or three additional items represents a large increase in the standard score or percentile rating. In other words, the individual who "gets stuck" on an item in a short test seems to be penalized unduly in his final rating. For this reason the *Classification Test for Industrial and Office Personnel*

contains 100 items, which are as many as appear in longer tests. Individuals at the college level will often answer correctly as many as 90 items and a small number of individuals (about 4 per cent) at this level will just about succeed in attempting every item in the maximum time allowed.

Type and Arrangement of Items. The 100 items are spiralled in series of five as follows: vocabulary, general information, arithmetic, general information, and analogics. There is thus a total of 40 general information items and 20 each of vocabulary, arithmetic, and analogies. The entire test is contained in a four-page booklet with the directions and practice problems on the first page and the test items on pages 2, 3, and 4. All of the items are of the multiple-choice type, with four alternates.

Time Limits. The time limit of the *Classification Test for Industrial and Office Personnel* has been kept to a minimum to make it practical to use in the employment situation. It is possible to use a time limit of either 10 or 15 minutes. The 15-minute time limit is recommended since the norms for this time limit are based on an appreciably larger number of cases than for the 10-minute period.

Standardization. Originally the test was administered in tentative form to over 3000 subjects. These included general college students, engineering college students, evening college students, high school students, nursing school applicants, clerical workers from typical manufacturing establishments, salesmen, and factory workers. The test went through two mimeographed versions and one printed version on an experimental basis before it was published in its final form.

Reliability and Equivalence of Forms. The odd-even reliability of the test, as corrected by the Spearman-Brown formula, is .94. Two forms of the test, A and B, are available. A correlation of .86 between the two forms was obtained when they were administered in A-B order to a group consisting of 90 academic high school students and 159 college students. A correlation of .85 was obtained for a group of 72 commercial high school students. Correlations of .80 and .82 were obtained for similar groups who took the tests in B-A order.

The differences in difficulty between the two forms are practically negligible and appear to approach the minimum that can be expected. For a group of 389 academic high school and college students, the difference was 1.34 raw score points. For a group of 67 commercial high school students the difference in difficulty was −.178 raw score points. The practice effect for these two groups was 4.76 and 2.42 respectively.

Validity. Validity coefficients obtained thus far are of two types: (1) correlations with other tests of intelligence, and (2) correlations with job or school performance. These validity coefficients are presented in Table 1. Since this test is short and does not cover the entire range of

mental ability, correlations between it and longer tests of intelligence are
not as high as are usually obtained between longer tests of intelligence.
As can be seen from Table 1, the test has demonstrated low but positive
validity in the industrial situation and somewhat better validity in the
commercial school situation. It would appear, however, that because
of its short time limit, the test is appropriate as part of a battery designed
for industrial or school use. It is of interest to note that a critical norm
of 40 was established in two validity studies. In the first of these the
test was used as part of a battery to select salesmen. It was found that
men scoring below 40 were difficult to train and inferior in sales perfor-
mance. In the second study it was found that men scoring below 40
were poor risks for the job of bus or street car operator.

Norms. The following norms are available: Adult (N = 1662);
general college (N = 946); engineering college (N = 113); evening college
(N = 293); high school (N = 383); nursing school applicants (N = 254);
clerical workers (N = 137); sales personnel (N = 225); factory workers
(N = 1494); general population (N = 6007).

Table 1

Validity Coefficients for the Classification Test for Industrial and Office Personnel

Group	N	r	Criterion
			Other Tests
College	191	.69	A.C.E., 1942 Edition
High School	83	.75	Otis S-A, Higher Forms B and D
Clerical Employees	100	.80	Otis S-A, Higher Forms A, B, and D
General Adult	149	.69	Otis S-A, Higher Form D
General Adult	105	.76	A.C.E., 1941 Edition
Nursing Applicants	254	.62	California Mental Maturity, Form A
Junior Clerks	44	.83	Otis S-A, Higher Form D
			School Course Grades
High School	123	.46	Business Information and Mathematics
High School	126	.27	Typing
High School	53	.37	Bookkeeping
High School	53	.38	Stenography
High School	46	.47	Office Production
High School	46	.56	Filing
High School	39	.46	Machine Calculation
			Job Performance
Maintenance Salesmen	45	.31	Ratings of sales performance
Maintenance Salesmen	45	.21	Total sales for two years
Heater Salesmen	79	.44	Sales ability (biserial r)
Junior Clerks	44	.49	Job rating
Junior Clerks	29	.43	Progress rating

Summary

A short test of mental ability has been described which, it is felt, is very appropriate for use in the industrial and vocational guidance situations. This test is the *Classification Test for Industrial and Office Personnel*, Forms A and B.

The distinguishing characteristics of this test are: (1) a short time limit; (2) a large number of items of approximately uniform difficulty, rather than a small number of items presented in order of increasing difficulty. It is believed that this sort of mental ability test is more suitable to the typical office or factory employment situation than the usual type of intelligence test.

At the present writing the test has been standardized on over 6000 subjects. The odd-even reliability is .94, and the correlation between alternate forms varies from .80 to .86. Differences in difficulty between the two forms are practically negligible. Norms are available for nine different industrial and school populations. Validities with other, longer, tests of mental ability range from .62 to .83. Validities with grades in commercial high school courses range from .27 to .56. Validities with various criteria of job performance range from .21 to .49.

References

1. Buros, O. K., Ed. *The nineteen forty mental measurements yearbook.* Arlington, Va.: Gryphon Press, 1945.
2. Otis, J. L. The Personnel Research Institute of Western Reserve University. *J. consult. Psychol.*, 1946, **10**, 131–135.
3. Otis, J. L., et al. *Classification test for industrial and office personnel* (Forms A and B). Cleveland, Ohio: Western Reserve University Press, 1947.

Abbreviated Job Evaluation Scales Developed on the Basis of "Internal" and "External" Criteria

David J. Chesler

Personnel Research Institute, Western Reserve University

In recent years much of the published material in the field of job evaluation which might properly be designated as "research" has been concerned with *abbreviated* job evaluation scales. Most of this work has been performed by Lawshe and various associates (4, 5, 6, 7). The writer (3) has also presented some findings on this topic. All of these studies utilized the Wherry-Doolittle selection method (8) to derive the abbreviated scales. The procedure has been to apply the Wherry-Doolittle process to the factors or "rating scale items" which comprise a job evaluation scale, and to identify the first three or four factors in the scale which contribute most to the ratings which jobs receive on the scale. The ratings predicted from these three or four factors are then compared with the ratings received on all of the original factors. The criterion is the original job evaluation scale from which the abbreviated scale was derived.

The present study has attempted to answer the question as to which three or four factors in a job evaluation scale would be identified if another job evaluation scale were used as the criterion. Such a criterion has been designated throughout this report as an "external "criterion, in contrast to the rating on the original manual, which may be designated as the "internal" criterion. Will similar abbreviated scales emerge when various job evaluation manuals constitute the external criteria? It is believed that a study of this sort offers a method of analyzing the differences between two job evaluation manuals. Specifically, it answers the question of what factors in one job evaluation system constitute the best measure of another system.

Method

Job raters in three industrial organizations rated independently descriptions and specifications for 35 "standard" salaried jobs on a "standard" job evaluation manual and on their own respective company manuals. The jobs, the standard manual, the company manuals, and the job analysts involved are the same as those reported in previous studies (2, 3).

Results and interpretation

Standard Manual Factors Identified with Internal Criterion. As reported previously (3), the Wherry-Doolittle selection method was applied to the standard manual factor ratings submitted by the raters in the three companies, with total rating on the standard manual as the (internal) criterion. With the internal criterion the first four factors identified with each of the three groups of raters were the same, although the order of identification was not the same.[1] These four factors were: "Work experience"; "character of supervision received"; "character of supervision given"; and "responsibility for confidential matters."

Table 1

Abbreviated Scales Derived from Standard Manual with External Criteria in Three Companies by Raters Who Rated the Standard Jobs on the Standard Manual and on Their Respective Company Manuals

Co. A		Co. B		Co. C	
Factor No.	$\bar{R}$	Factor No.	$\bar{R}$	Factor No.	$\bar{R}$
5	.839	6	.854	5	.905
2	.921	9	.926	4	.959
6	.941	10	.945	2	.969
10	.956	8	.956	11	.974
		11	.958	8	.977
				7	.978
				10	.979
				12	.979
				6	.978

Key to Factor Numbers: 2. Essential knowledge and training; 4. Character of supervision received; 5. Character of supervision given; 6. Number supervised; 7. Responsibility for funds, securities, and other valuables; 8. Responsibility for confidential matters; 9. Responsibility for getting along with others; 10. Responsibility for accuracy—effect of errors; 11. Pressure of work; and 12. Unusual working conditions.

Standard Manual Factors Identified with External Criterion. The procedure followed in the present study was to apply the Wherry-Doolittle selection process to the factors of the standard manual, with total ratings on a company manual as the (external) criterion. The results are summarized in Table 1.

Since comparisons of abbreviated scales have previously (3) been made on the basis of the first four factors identified, we need concern ourselves in Table 1 only with the first four factors identified in each

[1] For a more complete discussion of these findings, see a previous study (3).

instance. The striking feature of the abbreviated scales that emerge with different external criteria is their dissimilarity—as contrasted with the striking similarity of the abbreviated scales that emerged with the same internal criterion (3). The number of times certain factors were identified among the first four factors for the three groups of raters may be summarized as follows:

Factor Number	Factor	No. Times Identified
2.	Essential knowledge and training	2
4.	Character of supervision received	1
5.	Character of supervision given	2
6.	Number supervised	2
8.	Responsibility for confidential matters	1
9.	Responsibility for getting along with others	1
10.	Responsibility for accuracy-effect of errors	2
11.	Pressure of work	1
	Total	12

Out of a possible total of twelve factors, eight appeared either once or twice. It is interesting that no single factor emerged three times, that is, once in each of the abbreviated scales derived with an external criterion. Of the eight factors, three ("character of supervision received," "character of supervision given," and "responsibility for confidential matters") were also identified in abbreviated scales derived with the internal criterion (3). It would appear that these three factors are important, not only in the standard manual, but also in some form or other in the manuals used in the three companies. The fact that two "supervisory" items were identified, not only with the internal criterion, but also with three different external criteria would indicate that in the standard and the company manuals factors concerned with supervision are very important.

It would seem that the results obtained with external criteria indicate primarily essential differences among the company manuals, as analyzed in terms of the standard manual factors. This may be contrasted with the results obtained with the same internal criterion—which indicate primarily differences among the raters (3).

Adequacy of Abbreviated Scales Derived from Standard Manual with External Criterion in Predicting External Criterion. As in the case of abbreviated scales derived with internal criteria (3), an analysis was made of the accuracy with which the abbreviated scales, derived from the standard manual with external criteria, predict the external criteria, that is, ratings on the company manuals.

The multiple regression equations for predicting total points on the company manuals from point ratings on the selected standard manual factors were computed and applied to the ratings given on the selected standard manual factors by the raters in the three companies. Three sets of predicted company manual ratings were thus obtained.

The actual classification plans of the three companies (see Table 2)[2] were used to study the comparative adequacies of the three abbreviated scales. The labor grades within each of the company plans are unequal and follow roughly a geometric rather than an arithmetic progression.

Table 3 shows the per cent of jobs in each instance which remained in the same labor grade, or which were displaced into another labor grade. In companies A, B, and C, respectively 88.5 per cent, 91.4 per cent, and 97.1 per cent of the jobs remained in the same labor grade or were displaced into a labor grade adjacent to that of the original classification. In all three companies some jobs were displaced by two or three labor grades.

Table 4 shows how ratings on the abbreviated scales derived with external criteria deviated by the point value of 0.5 labor grade, 1.0 labor grade, or more than 1.0 labor grade from total ratings on the original (company) manuals. In the three companies respectively 31.4 per cent, 54.3 per cent, and 68.6 per cent of the predicted ratings deviated from the original ratings by the point value of 0.5 labor grade or less. Similarly 65.6 per cent, 85.7 per cent, and 88.6 per cent of the predicted ratings deviated from the original ratings by the point value of 1.0 labor grade or less. In other words, for three companies respectively 34.4 per cent, 14.3 per cent, and 11.4 per cent of the predicted ratings deviated from the original ratings by a point value greater than one labor grade.

Adequacy of Abbreviated Scales Derived from Standard Manual with Internal Criterion in Predicting External Criterion. The first four factors of the standard manual consistently identified by the Wherry-Doolittle selection process with total rating on the standard manual as the (internal) criterion were factors 1, 4, 5, and 8, that is "work experience," "character of supervision received," "character of supervision given," and "responsibility for confidential matters" (3). These factors might be described as the primary factors of the standard manual because, when weighted properly, they are the "best measure" of total ratings on the standard manual. It is of interest to know how well this best measure of a manual measures total ratings on other manuals.

The specific question to be answered here is how well do the abbreviated scales derived from the standard manual with total ratings on the standard manual as the (internal) criterion predict the external criterion, that is, company manual ratings.

The multiple regression equations for predicting company manual ratings from factors 1, 4, 5, and 8 of the standard manual were computed and applied to the ratings given on these factors by the raters in the three companies. Three sets of predicted company manual ratings were thus obtained.

Again, the actual classification plans of the three companies (see Table 2) were used to study the comparative adequacies of the abbreviated scales. Table 5 shows the per cent of jobs in each instance which remained in the same labor grade or which were displaced into labor grades one, two, or more grades removed from that of the original classification. In companies A, B, and C, respectively 68.5 per cent, 94.2 per cent, and 91.4 per cent of the jobs remained in the same labor grade or were displaced into a labor grade adjacent to that of the original classification.

Table 6 shows how predicted company manual ratings based on the abbreviated scales derived with internal criteria deviated by the point value of 0.5, 1.0, or more than 1.0 labor grade from total ratings on the original (company manual) scales. In the three instances 11.5 per cent, 40.0 per cent, and 60.0 per cent of the predicted ratings deviated from the original ratings by the point value of 0.5 labor grade or less; and 42.8 per cent, 71.4 per cent, and 91.4 per cent of the predicted ratings deviated from the original ratings by the point value of 1.0 labor grade or less.

Comparison of All Abbreviated Scales Derived from Standard Manual. In both the present and in a previous study (3) abbreviated scales have been derived from a standard manual, with an internal criterion (standard manual total rating) and with an external criterion (company manual total rating). Table 7 summarizes the data required to form an opinion as to the relative adequacies of these abbreviated scales in predicting either the internal or external criterion.

In terms of the multiple coefficient of correlation and the index of forecasting efficiency the adequacy of prediction is clearly in the hierarchy:

1. Abbreviated scales derived with internal criterion and used to predict the internal criterion.

2. Abbreviated scales derived with external criterion and used to predict the external criterion.

3. Abbreviated scales derived with internal criterion and used to predict the external criterion.

Table 7*

Adequacy of Abbreviated Scales Derived from Standard Manual with Internal
Criterion (Standard Manual Ratings) and External Criterion
(Company Manual Ratings)

	Co.	Derived with internal criterion; used to predict internal criterion	Derived with external criterion; used to predict external criterion	Derived with internal criterion; used to predict external criterion
Multiple coefficient of corre-	A	.98	.96	.91
lation (R)	B	.98	.96	.88
	C	.99	.98	.97
Index of forecasting efficiency	A	.79	.72	.59
(E)	B	.81	.72	.52
	C	.85	.78	.75
% jobs remaining in same, or	A	100.	88.5	68.5
displaced into adjacent, labor	B	100.	91.4	94.2
grade	C	100.	97.1	91.4
% predicted ratings deviating	A	94.2	65.6	42.8
value of 1.0 labor grade or less	B	94.2	85.7	71.4
from original ratings	C	97.1	88.6	91.4

* See footnote 2.

This hierarchy is apparent from the fact that all R's and E's decrease as one reads across each row.

In terms of the percentage of jobs remaining in the same or in being displaced into an adjacent labor grade, this hierarchy holds for companies A and C, but not for Co. B. However, in the case of Co. B the discrepancy is due to a difference of only one job.

In terms of the percentage of predicted ratings deviating by the value of one labor grade or less from the original ratings, the hierarchy holds for companies A and B, but not for Co. C. However, here again the discrepancy is due to a difference of only one job.

Summary

1. The basic methodological feature of the present study was to have raters in three companies evaluate a standard set of descriptions and specifications for 35 representative salaried jobs on a standard job evaluation manual and on their own respective company manuals.

2. The Wherry-Doolittle selection method was applied to the standard manual factor ratings submitted by the raters in each company, with

total rating on the standard manual as the (internal) criterion. The first four factors identified were the same for each group of raters, although the order of identification was not the same (3). These results indicate primarily differences among raters.

3. The Wherry-Doolittle selection method was again applied to the standard manual factor ratings submitted by the raters in each company, but with total ratings on the respective company manuals as the (external) criterion. Out of a possible total of twelve factors, eight were identified among the first four for all three groups of raters. The striking feature of the abbreviated scales derived with external criteria is their dissimilarity—as contrasted with the striking similarity of the abbreviated scales that emerged with the same internal criterion. These results indicate primarily differences among the company manuals, as analyzed in terms of the standard manual factors.

4. An analysis of the adequacy of the abbreviated scales derived from the standard manual with internal and external criteria in predicting the internal and external criteria indicates in general the following hierarchy of accuracy of prediction with abbreviated scales:

a. Derived with internal criterion and used to predict the internal criterion.

b. Derived with external criterion and used to predict the external criterion.

c. Derived with internal criterion and used to predict the external criterion.

References

1. Benge, E. J., Burk, S. L. H., and Hay, E. N. *Manual of job evaluation.* New York: Harper & Brothers, 1941.
2. Chesler, D. J. Reliability and comparability of different job evaluation systems, *J. appl. Psychol.,* 1948, 32, 465–475.
3. ———. Reliability of abbreviated job evaluation scales. *J. appl. Psychol.,* 1948, 32. 622–628.
4. Lawshe, C. H., Jr. Studies in job evaluation: II. The adequacy of abbreviated point ratings for hourly-paid jobs in three industrial plants. *J. appl. Psychol.,* 1945, 29, 177–184.
5. ———, and Alessi, S. L. Studies in job evaluation: IV. Analysis of another point rating scale for hourly-paid jobs and the adequacy of an abbreviated scale. *J. appl. Psychol.,* 1946, 30, 310–319.
6. ———, and Maleski, A. A. Studies in job evaluation. 3. An analysis of point ratings for salary paid jobs in an industrial plant. *J. appl. Psychol.,* 1946, 30, 117–128.
7. ———, and Wilson, R. F. Studies in job evaluation. 5. An analysis of the factor comparison system as it functions in a paper mill. *J. appl. Psychol.,* 1946, 30, 426–434.
8. Stead, W. H., Shartle, C. L., and Associates. *Occupational counseling techniques.* New York: American Book Co., 1940.

Studies in Job Evaluation: 8. The Reliability of an Abbreviated Job Evaluation System

C. H. Lawshe and Patrick C. Farbro

Occupational Research Center, Purdue University

Several systems for evaluating jobs have been developed. Of these much has been written and considerable experimentation has been carried on because the setting of wage rates is one of the most important managerial functions. The great majority of these systems arrive at their goal—the systematic pricing of jobs—by breaking the jobs into their various elements or components. The number of elements on which jobs have been rated varies from system to system. Using a scaling method of some sort, the rater assigns degrees of each component to each job, the various degrees are weighted and total point values are converted into wage rates.

Previous Studies. As a result of a series of studies by the senior author and others (1, 2, 3, 4, 5, 6), an abbreviated system of job evaluation has been developed and reported. When forty job descriptions were submitted to two groups of independent raters, one of which applied the NEMA system and the other used this system, a correlation of .90 between the two was obtained (7). Lawshe and Wilson (6) have shown in a previous study that the abbreviated system of four items is more reliable (.98 for five raters) than the NEMA system (.94 for five raters). However, since their data were gathered by sending job description by mail to the cooperating analysts, the question of functional reliability in the practical situation remains unanswered.

Purpose of this Study. The primary purpose of this study was to determine the reliability or consistency with which raters, all from the same plant, evaluate jobs in that plant by means of this simplified system.

More specifically, the purposes of this study are: (1) to compare reliability coefficients obtained in Lawshe and Wilson's study of hypothetical jobs with reliability coefficients in an operating plant; (2) to compare independent ratings made by the evaluation committee with ratings adjusted through conference discussion; and (3) to examine rating differences between labor committee members and management committee members.

Procedure

The Abbreviated Job Evaluation System. The system of job evaluation used in this study is that developed by Lawshe. The system provides for the rating of jobs on four scales: "General Schooling," "Learning Period," "Working Conditions," and "Job Hazards."

The Job Evaluation Committee. The committee used in evaluating the forty-three jobs in this study consisted of five members. Two of the members were employees belonging to the union. Management representatives on the job evaluation committee included the production manager and the secretary of the company. The fifth member of the committee was the production superintendent during the time production jobs were being evaluated and the maintenance superintendent while maintenance jobs were being evaluated.

Rating the Jobs. The actual procedure of rating the jobs consisted of several phases which were preceded by standard job description preparation. After a general orientation, each committee member was furnsihed a set of forty-three 3″ by 5″ white cards on which had been typewritten each of the forty-three job titles. The committee was then instructed to consider only the "General Schooling" required for performing each job and on that basis to place the cards in rank order from the job requiring the greatest amount of schooling to the job requiring the least amount of schooling for successful performance on the job. On completion of this task, a set of six colored cards representing each of the six degrees of the "General Schooling" scale was given each committee member. Members were then instructed to insert the colored cards in their stack of white cards at the places most logical for the breaks. Thus the degrees of the "General Schooling" scale were assigned. In similar manner, "Learning Period," "Working Conditions," and "Job Hazards" scales were employed in rating each job.

From these cards a summary page showing the degrees assigned each job by each committee member was prepared and anchor jobs, those on which at least four of the members initially agreed, were identified. The committee was then again assembled and by using the anchor jobs as reference points, members discussed and adjusted the ratings for those jobs on which there was disagreement. It is important, however, that the initial ratings were made without committee discussion.

Results

The Obtained "one against one" Reliability Coefficients. Shown in Table 1 in the second column are the obtained reliability coefficients for each of the items of the abbreviated system and for total points as jobs

were evaluated in this study. The figures shown in this column are the averages of the coefficients obtained by correlating initial ratings of each rater with initial ratings of every other rater.[1] The figures shown in column two are the most likely correlations between the ratings of one rater and the ratings of one other rater. For convenience and for comparison with the previous study by Lawshe and Wilson (6), these have been called the "one against one" reliabilities.

Table 1

Reliability Coefficients for Total Point Ratings and for the Component Scale Ratings in the Lawshe-Wilson Study and in This Study

Item	"One against one" Reliability		"Five against five" Reliability	
	Lawshe-Wilson	This Study	Lawshe-Wilson	This Study
Total Points	.89	.91	.98	.98
Learning Period	.86	.84	.97	.96
General Schooling	.79	.84	.95	.96
Working Conditions	.61	.73	.89	.93
Job Hazards	.51	.54	.84	.86

The "five against five" Reliability Coefficients. Even though the reliability coefficients shown in column two are those actually obtained, they are inadequate estimates of the true reliability of pooled ratings of members of the committee. As was mentioned before, the "one against one" reliabilities are the best estimate of reliability of the ratings of one rater as compared with those of one other rater. Since five raters were involved in the rating of each job, the reliabilities in the second column were "stepped up" by use of the Spearman-Brown formula to estimate the reliabilities of the pooled ratings of all five of the job evaluation committee members. These "stepped up" ratings are shown in Table 1 in the fourth column.

It is not advocated that these coefficients of reliability be accepted as absolute, but merely that they are estimates of the ture reliabilities of the abbreviated job evaluation system. The results presented should be qualified in view of one's own evaluation of the assumptions involved in such a procedure.

It is evident from column four that reliability coefficients of the magnitude found are definitely high enough for purposes for which the system was designed. As will be noted, "five against five" reliability coefficients

[1] Correlations were found between the following patterns of pairs of raters: A-B, A-C, A-D, A-E, A-F, B-C, B-D, B-E, B-F, C-E, C-F, D-E, D-F, E-F. Obtained reliability coefficients were then averaged by transformation to Fisher Z-values (9).

for the four scales range from .86 (Job Hazards) to .96 (Learning Period), with all but one scale, "Job Hazards," above .90. Agreement among raters as evidenced by a reliability coefficient of .98 for total points definitely indicates high enough reliability for most practical purposes.

Comparison of Lawshe-Wilson Study and This Study. The first item of interest in comparing the data from the two studies in Table 1 is the close agreement of the reliabilities found for total point ratings (.89 and .91 for "one against one" reliabilities and .98 and .98 for "five against five" reliabilities).

The single items found most reliable in this study are those of the "Skill Demands" factor (Learning Period and General Schooling) and are the same as those found most reliable in the previous study.

The next most reliable items in this study (Working Conditions) has the same rank position in the Lawshe-Wilson study. The least reliable scale of the abbreviated system (Job Hazards) was found in the same rank-position in both studies.

It is interesting to note that the rank order of magnitude of the reliability coefficients is essentially the same in both studies. It appears that the estimation of the reliabilities in the Lawshe-Wilson study were a conservative estimate of the reliability of the abbreviated system when employed in an actual industrial situation. This is easily understood since in the Lawshe-Wilson study the several raters were from different plants in scattered geographical locations and used only job titles and descriptions in evaluating the jobs, while in this study five raters were evaluating definite jobs in a plant with which each was familiar.

Comparison of Management and Labor Ratings. In comparing the reliability of labor and management committee members, the first item of interest is the consistency of the findings as shown in Table 2. The correlation coefficients representing reliability or agreement between the labor committee members are consistently lower than those representing agreement between two management members. The "one against one" reliability coefficients for labor union committee members range from .37 (Working Conditions) to .83 (Total Points), while for management members they range from .80 (Job Hazards) to .94 (Total Points).

Considering agreement between two labor union committee members and two management representatives[2] "one against one" reliability coefficients range from .66 (Job Hazards) to .86 (Total Points). These reliability coefficients, it will be noted from Table 2, fall between those of the labor members which are lowest and those of management representatives which are highest.

[2] Mean of obtained correlations between each management member and each labor member was derived by transformation to Fisher Z-values.

Table 2

Coefficients of Reliability for Two Labor and Two Management Job
Evaluation Committee Members

Item	Labor-Labor	Mgmt-Mgmt	Labor-Mgmt
Total Points	.83	.94	.86
Learning Period	.73	.86	.80
General Schooling	.71	.92	.77
Working Conditions	.37	.90	.71
Job Hazards	.68	.80	.66

Comparison of Initial Ratings with Adjusted Ratings. As was previously mentioned two sets of ratings were available for each job title—initial ratings, independently assigned by the raters for each of the various scales for each job title, and adjusted ratings, those resulting from conference discussion. The mode of the adjusted ratings was the point value actually used as the basis for the wage structure in the plant. The mean or average points assigned by each were used in the Lawshe-Wilson study since it was impossible to assemble the various raters in a conference for the purpose of adjusting ratings. For this reason it was considered advisable to obtain a measure of relationship between the mean initial ratings and the mode of adjusted ratings. In Table 3 the correlation is shown to be .97 for Total Points, and to range from .83 (Job Hazards) to .94 (General Schooling) for the component scales of the abbreviated system. These values are probably large enough to support the hypothesis that conclusions based upon mean independent ratings are valid for plant situations in which majority decisions are reached.

Similarly, it seemed desirable to investigate separately the relationship between initial ratings as made by management, labor, and maintenance and production superintendents and the mode of adjusted

Table 3

Coefficients of Correlation Between Mean of Initial Ratings and
Mode of Adjusted Ratings

Item	r
Total Points	.97
Learning Period	.92
General Schooling	.94
Working Conditions	.86
Job Hazards	.83

ratings.   Table 4 shows these relationships.   Coefficients of correlation between the mean of management representatives' initial ratings and the mode of adjusted ratings were found to be consistently larger (ranging from .73 on "Job Hazards" scale to .97 for Total Points) than those of labor union members (.65 to .93 on the same scales).   The rank order of magnitude of the coefficients is the same for both management and labor union committee members.

Table 4

Obtained Coefficients of Correlation Between Initial Ratings for Management, Labor, and Superintendents, and Mode of Adjusted Ratings

Item	Mgmt	Superintendents		Labor
		Maint.	Prod.	
Total Points	.97	.97	.97	.93
Learning Period	.95	.96	.96	.92
General Schooling	.93	.93	.88	.86
Working Conditions	.86	.85	.90	.69
Job Hazards	.73	.77	.89	.65

Also shown in Table 4 are the correlations between the maintenance and production superintendents' initial ratings and the mode of adjusted ratings.   The magnitude of these coefficients (ranging from .77 on "Job Hazards" scale to .97 for Total Points) is higher than those of the labor union committee members.   They are also larger than those of the management committee members on all but the "General Schooling" scale.

In considering the change from initial ratings to adjusted ratings, it was also deemed advisable to examine actual point value changes.   This was accomplished by tabulating each rater's actual point change from his initial ratings to his adjusted ratings.   The point value for each item and total points assigned by each rater for each job was considered in this analysis.   For example, on job number 1, Rater A initially rated the job as being worth 130 points on the "Learning Period" scale.   During the conference discussions his rating was changed to 150 points; thus a +20 was tabulated.   This procedure was followed throughout.

Shown in Table 5 is the mean gross change per job by raters from initial point ratings to adjusted point ratings.   These point changes were derived as above by adding point value changes disregarding algebraic signs.   From Table 5 a general trend may be seen.   Raters "E" and "F", both labor members, have the greatest average gross change from initial to adjusted ratings while Rater "C", the maintenance superintendent,

Table 5

Mean Gross Change per Job of Raters from Initial Point Ratings
to Adjusted Point Ratings

Rater	N	Total Points	Learning Period	General Schooling	Working Conditions	Job Hazards
A (Mgmt)	43	10.4	6.9	3.1	1.3	.7
B (Mgmt)	43	10.1	4.4	5.0	.9	.9
C (Sup-Maint)	16	1.8	1.2	0.0	.4	.2
D (Sup-Prod)	27	10.2	3.8	5.2	.2	1.2
E (Labor)	43	18.6	8.7	12.4	1.1	.9
F (Labor)	43	18.3	12.0	5.7	1.9	.8

changed his initial ratings the least. Raters "A" and "B", the two
management committee members, changed less (10.4 and 10.1 average
points per job, respectively) from initial ratings to adjusted ratings than
did the labor committee members (18.6 and 18.3 average points per
job for Raters "E" and "F", respectively).

In Table 6 the average net change per job by raters from initial point
ratings to adjusted point ratings is shown. These values were obtained
in the same manner as described above except that algebraic signs were
considered. In general the trend shown in Table 6 is that Raters "A"
and "B", the two management raters, and Raters "C" and "D", the
production and maintenance superintendents, initially tended to under-
rate the jobs except in relation to the "Working Conditions" scale;
therefore, they had to increase point ratings in the conferences, while
Raters "E" and "F", both labor union members, over-rated jobs on the
"General Schooling" and "Working Conditions" scales but under-rated
on the "Learning Period" and "Job Hazards" scales.

It is interesting to note that the mean of the initial points as conceived

Table 6

Average Net Change per Job of Raters from Initial Point Ratings
to Adjusted Point Ratings

Rater	N	Total Points	Learning Period	General Schooling	Working Conditions	Job Hazards
A (Mgmt)	43	+6.9	+5.3	+1.8	−1.3	+ .6
B (Mgmt)	43	+2.6	+ .1	+2.6	− .9	+ .8
C (Sup-Maint)	16	+1.3	+1.2	0	+ .4	− .3
D (Sup-Prod)	27	+7.9	+3.8	+3.0	+ .1	+1.2
E (Labor)	43	−2.0	+5.6	−6.9	− .9	+ .2
F (Labor)	43	+8.3	+9.3	− .3	−1.3	+ .6

by the evaluation committee is 250.30 while after conference discussion the mean of adjusted ratings is 254.63. In analyzing this difference of 4.33 points, a critical ration of 2.47 was found, indicating the difference to be significant at the 2 per cent level of confidence.

Summary and Conclusions

Job evaluation data for forty-three jobs from a manufacturing plant using an abbreviated evaluation system were analyzed. The job evaluation committee, including two management members, two employees affiliated with the labor union active in the plant, and the maintenance superintendent or production superintendent when maintenance or production jobs were being considered, evaluated each job on the four items of the abbreviated system.

Reliability coefficients for the total point ratings and for the individual scales were obtained by correlating ratings given each job title on the basis of each of the four factors of the evaluation system. Correlations were found between the ratings of each rater as paired with every other rater and these obtained coefficients were averaged after transformation to Fisher Z-values. These average intercorrelations were then stepped-up using the Spearman-Brown formula to obtain the estimated reliability of the ratings of the five-member committee.

Comparison was made with a previously published study by Lawshe and Wilson which employed the abbreviated evaluation system. Analysis was also made comparing independent ratings with ratings adjusted after conference discussion. Differences in agreement or consistency of ratings with which labor and management committee members rate jobs were also explored.

The following conclusions are supported:

1. The abbreviated system demonstrates reliability sufficiently high for most practical purposes (.98 for five raters).

2. Comparisons with the Lawshe-Wilson study shows the same rank-order pattern of reliabilities for total points and the individual scales.

3. In comparing relative reliabilities of management and labor committee members, those of management were found consistently higher (ranging from .80 to .94); than those of labor (ranging from .37 to .83). The magnitude of reliability or agreement of average intercorrelations of the management and labor union committee members combined (range from .66 to .86 for one rater) falls between those of labor union members which are of lowest magnitude and those of management representatives which are of highest magnitude.

4. High correlation was found between mean initial ratings (those assigned independently) and the mode of adjusted ratings (adjusted during conference discussion) for Total Points (.97) and also for the component items of the system (ranging from .83 to .92).

5. Initial ratings as conceived by management, labor, and the superintendents separately when compared with the mode of adjusted ratings showed the superintendents to initially rate jobs more accurately as evidenced by correlation with adjusted ratings than did management committee members or labor union committee members. Management members' initial ratings agreed more closely with final ratings than did labor union members' initial ratings. The same relationship was found in analyzing actual point changes from initial to mode of adjusted ratings.

6. Throughout this analysis the Skill Demands factor as measured by "Learning Period" and "General Schooling" items was found the most stable as evidenced by the fact that average intercorrelations for these two scales were largest; that agreement between management and labor members on these two scales was greater than on other scales; and that correlations between initial and adjusted ratings on these two scales was higher than on scales of "Job Characteristics" factor.

References

1. Lawshe, C. H., Jr., and Satter, G. A. Studies in job evaluation. 1. Factor analysis of point ratings for hourly paid jobs in three industrial plants. *J. appl. Psychol.*, 1944, 28, 189–198.

2. Lawshe, C. H., Jr. Studies in job evaluation. 2. The adequacy of abbreviated point ratings for hourly-paid jobs in three industrial plants. *J. appl. Psychol.*, 1945, 29, 177–184.

3. Lawshe, C. H., Jr. Studies in job evaluation. 3. An analysis of point ratings for salary paid jobs in an industrial plant. *J. appl. Psychol.*, 1945, 30, 117–128.

4. Lawshe, C. H., Jr., and Alessi, S. L. Studies in job evaluation. 4. Analysis of another point rating scale for hourly-paid jobs and the adequacy of an abbreviated scale. *J. appl. Psychol.*, 1946, 30, 310–319.

5. Lawshe, C. H., Jr., and Wilson, R. F. Studies in job evaluation. 5. An analysis of the factor comparison system as it functions in a paper mill. *J. appl. Psychol.*, 1946, 35, 426–434.

6. Lawshe, C. H., Jr., and Wilson, R. F. Studies in job evaluation. 6. The reliability of two point rating systems. *J. appl. Psychol.*, 1947, 31, 355–365.

7. Lawshe, C. H., Jr., Dudek, Edmund E., and Wilson, R. F. Studies in job evaluation. 7. A factor analysis of two point rating methods of job evaluation. *J. appl. Psychol.*, 1948, 32, 118–129.

8. Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill Book Co., 1940.

9. Snedecor, G. W. *Statistical methods.* Ames, Iowa: Iowa State College Press, 1946.

Odor Selection, Preferences and Identification

Bernard Locke and Charles H. Grimm

Brooklyn, N. Y.

In light of the fact that many millions of dollars are spent annually in the purchase of aromatic products it is extremely surprising that so little work has been done in any systematic fashion to evaluate some of the factors which lead an individual to select a particular aromatic compound for purchase. It is the purpose of this paper to explore, in a preliminary fashion, several of the factors which might play a part in such selection.

The broad elements to be dealt with in this research include: 1. The ability to differentiate between "expensive" and "inexpensive" odors. 2. The relationship between subjective concepts of costliness and "pleasantness" or "unpleasantness" of a perfume compound. 3. The ability to recognize some of the more common floral odors.

The 69 female subjects used were a select rather than a cross section sampling in that they were students in an advanced collegiate course in psychology and our interpretations of the results will, therefore, take this into consideration. The average age of the group was 24.7 years with a range from 19 to 50 years. The length of time that these individuals had been using perfumes ranged from one to twenty-five years with a mean of 7.2 years.

Experiment 1. The Ability to Differentiate Between "Expensive" and "Inexpensive" Odors

A search of the psychological literature for the past five years reveals only one experimental exploration of the ability of individuals to differentiate between expensive and inexpensive perfumes. In this experiment G. M. Jewett[1] employed three pairs of perfumes each containing an inexpensive member (50¢ an ounce) and an expensive one ($8.00 to $16.00 per ounce). His subjects were asked to compare them as to general "desirability" or affect and "lasting quality" purely on the basis of the smell stimulus. Jewett concluded from his data that in both respects the inexpensive perfumes produced substantially the same results as the expensive.

---

[1] Jewett, G. M. A note on the relation between subjective estimates of the desirability and the lasting quality of certain perfumes and their cost. *J. gen. Psychol.*, 1945, **33**, 285–290.

In the present experiment the 69 subjects were individually given perfumers' blotters that had been dipped into standard strength samples (16 oz. of oil to 128 oz. of alcohol) of eight perfumes and asked to indicate on a check sheet whether they thought the perfume to be an expensive or inexpensive one and at the same time whether they thought it a pleasant or unpleasant one. A description of the perfume oils, odor types and their costs is as follows.

Each of the oils has been found to be commercially acceptable and has been in use for a period of years. The average cost of the inexpensive oils (Numbers 1, 3, 5 and 7) is $5.00 per pound and the average cost of the expensive compounds (Numbers 2, 4, 6 and 8) is $60.00 per pound. The floral odors used were selected for their high fidelity in reproducing the actual floral note demonstrated in many years of use. Odor No. 1. A heavy sweet, balsamic, amber type; 2. A subtle chypre-floral, French, modern bouquet; 3. A modern, sweet, resin, aldehyde-chypre type; 4. A modern, floral-spice, fantasy type; 5. A sweet, modern, trefle, "outdoor" type; 6. A sophisticated, aldehyde-floral, fantasy type; 7. A modern, aldehyde, French type; and 8. A heavy, sweet, balsamic, amber type.

Table 1

Subjective Estimates of Cost of Eight Perfume Samples

Note: Items Marked with a * Are the "Expensive" Compounds

Perfume No.	Inexpensive	Expensive	Per Cent of Correct Responses
1	49	20	71
2*	26	43	62
3	41	28	59
4*	44	25	36
5	41	28	59
6*	36	33	48
7	44	25	64
8*	39	30	43

Table 1 presents the selections. The range of correct estimations of cost runs from 36 per cent to 71 per cent. If the responses for all eight odors are averaged the mean percentage of correct responses is 55, or just slightly better than if the selections had been made purely by chance. However, if we consider the accuracy of the judgments as regards the expensive and inexpensive odors separately we find that 63.3 per cent of the subjects made accurate choices of the inexpensive odors as compared to 47.25 per cent correct choices for the expensive odors. The computed critical ratio is 2.56 indicating that the difference is significant at the 2 per cent level but not at the 1 per cent level.

If one considers the direction of the errors made it is found that in

38 per cent of the estimations inexpensive perfume compounds were classed as "expensive" while 53 per cent of the estimations of the expensive compounds categorized them as being "inexpensive." Thus, we note a distinct tendency to minimize rather than to exaggerate the "values" of the odor samplings.

The mean number of correct identifications as to relative costliness of the eight perfume samples was 4.4. Not one of the 69 individuals was able to classify all eight correctly nor did any individual fail to make a single correct choice.

In order to determine whether length of use plays any part in developing skill in differentiation between expensive and inexpensive odors the group was divided into those who had used perfume from 0 to 5 years (N = 32) and those who had been using it for 6 or more years (N = 37). A comparison of the number of correct selections of the members of these two groups reveals that there is no demonstrable improvement in ability to differentiate between the expensive and inexpensive odors with increasing numbers of years of perfume usage. This is best demonstrated by the fact that the average number of appropriate selections for both of the groups is exactly identical, namely, 4.4 correct.

In order to evaluate the role of frequency of use as opposed to length of use of perfume in developing the ability to differentiate between expensive and inexpensive perfumes the subjects were asked to indicate the frequency with which they used perfumes. This was done on a four point check list which was made up of the following steps: Frequently, Occasionally, Rarely, Not at all. The need for such an evaluation is best illustrated by the response of the oldest member of the group who, in reporting the number of years that she had used perfume, replied, "Once a year for twenty-five years." Because of the small size of the experimental group the one subject who fell in the "not at all" category, and who, incidentally, made 5 correct selections, has been thrown into the "rarely" group. The results indicate that for the present experimental sample there is no measurable difference in ability to discriminate expensive from inexpensive perfumes among individuals who use perfumes frequently, occasionally or rarely, the mean number of correct choices being 4.3, 4.5 and 4.5 respectively.

Experiment 2.   Relationship Between Subjective Concepts of Costliness and "Pleasantness" or "Unpleasantness" of a Perfume Compound

Since it is fairly common experience that with some individuals commodities can be "costly" and still "unpleasant" and vice versa it was decided to explore the frequency with which such variations occurred.

At the time that each of the subjects determined whether a sample was expensive or inexpensive she was also asked to indicate whether the odor was pleasing or unpleasant to her.　Table 2 presents the frequency with which each of the eight odors used in Experiment 1 was designated with the apparently contradictory adjectives "Inexpensive and Pleasant" or "Expensive and Unpleasant."　From this table we note that a considerable amount of disagreement exists between the individual's evaluation of the cost of each of the perfumes and its pleasantness.　This difference actually constitutes an average of 31.5 per cent or, virtually, one-third of the total number of comparisons made.　When the discrepancies for the "expensive" and "inexpensive" groups of perfumes are compared no difference is found.　The mean percentage of differences is 31.8 per cent for the inexpensive odors and 31.3 per cent for the expensive group. While there was a slightly greater tendency to attribute unpleasantness to odors thought to be costly than to consider as pleasant those compounds which were thought to be inexpensive, this difference is not sufficiently great to be significant.

Table 2

Differences in Subjective Concepts of Costliness and Pleasantness or
Unpleasantness of 8 Perfume Compounds

Note: Those Perfumes Marked with a * Are Expensive

Perfume No.	Inexpensive-Pleasant	Expensive-Unpleasant	Total Disagreement	Total Agreement
1	13	6	19	50
2*	15	11	26	43
3	10	14	24	45
4*	14	9	23	46
5	12	14	26	43
6*	9	10	19	50
7	5	14	19	50
8*	3	16	19	50

In considering the number of instances in which there was disagreement between the concepts of costliness and pleasantness for each of the individuals we learn again of the disagreement in attitudes between cost and pleasantness.　The mean number of disagreements for each of the individuals in terms of pairing inexpensiveness and pleasantness is 1.2 and the mean for the expensive-unpleasant pair is 1.5.　In one instance where the subject classed all of the perfumes as expensive, she also considered them all as being unpleasant.

Experiment 3.   An Investigation of the Ability of a Group of Subjects
to Recognize Some of the More Common Floral Odors

This section of the research was intended to examine the ability of the
same experimental group to identify some of the more common floral
odors.   The eight odors used were Lilac, Gardenia, Carnation, Rose,
Pine, Jasmin, Lily of the Valley and Geranium presented in that order.
Each subject was permitted to smell each of the odors on perfumers'
blotters after having been told that each of the odors that she would now
smell was that of a flower and that she was to identify it by name.   Table
3 shows the number of correct identifications of each of the eight odors.

Table 3

Correct Identifications of Eight Floral Odors

Floral Odor	Number of Correct Identifications (N = 69)	Correct Identification in Percentages
Lilac	24	35
Gardenia	23	33
Carnation	21	30
Rose	17	25
Pine	28	41
Jasmin	1	1
Lily of the Valley	16	23
Geranium	0	0

Examination of Table 3 shows that the range of correct identifications
of the floral odors used ranges from 0 (for gernaium) to 28 (for pine) or
from 0 to 41 per cent of correct identifications.   If one averages the
correct responses for all of the eight odors the resultant percentage of
correct responses is 23.5.   The apparent order of difficulty in identifica-
tion ranging from most difficult to least difficult is Gernaium, Jasmin, Lily
of the Valley, Rose, Carnation, Gardenia, Lilac and Pine.   While it is
somewhat surprising that so much difficulty was evidenced in identifying
the various odors it is particularly interesting that the rose which is so
common and popular to our culture caused so much difficulty in recog-
nition with only one out of every four subjects being able to identify it
correctly.

Table 4 presents the findings for the number of correct identifications
by each member of our experimental group.   This table reveals that 12
per cent of our subjects were unable to identify even one of the floral
odors used and, similarly, there was no individual who identified more
than four of the eight floral odors that we used.

Table 4

Correct Identifications of the Series of Eight Floral Odors

Number of Correct Identifications	Number of Individuals (N = 69)	Per Cent of Total Group
0	8	12
1	22	31
2	17	25
3	16	23
4	6	9
5	0	0
6	0	0
7	0	0
8	0	0
	Mean = 1.8	100

To illustrate the wide deviations in identification made by members of the group Table 5 presents the identities attributed to our samples of Rose and Carnation.

In order to determine the effect of knowledge of the identity of the floral odors under investigation upon the accuracy of the identifications one-half of the experimental group (35 subjects) was asked to repeat this portion of the experiment but this time they were given the names of the odors presented in random order.   Table 6 presents a comparison

Table 5

Identifications of Rose and Carnation Samples made by the 69 Subjects

Rose		Carnation	
Don't Know	27	Don't Know	24
Rose	17	Carnation	21
Lily, Lily of the Valley, Easter Lily	5	Gardenia	5
Gardenia	4	Geranium	4
Lilac	3	Jasmin	3
Sweet Pea	3	Spice	3
Jasmin	2	Rose	2
Bouquet	2	Orange Blossom	2
Cold Cream	1	Chrysanthemum	1
Baby's Breath	1	Mint	1
Orange	1	Lavender	1
Lemon Verbena	1	Musk Blossom	1
Geranium	1	Clover	1
Carnation	1		
	69		69

Table 6

Comparison of Accuracy of Identification of Floral Odors with and without Knowledge of Their Identities

Floral Odor	Correct Responses with Knowledge of Identities	Correct Responses without Knowledge of Identities
Lilac	57%	35%
Gardenia	46%	33%
Carnation	54%	30%
Rose	23%	25%
Pine	94%	41%
Jasmin	20%	1%
Lily of the Valley	40%	23%
Geranium	20%	0%
	Mean 44%	Mean 23%

of the findings for this group and the original group. Examination of this table reveals a rather marked improvement in identifications in all of the odors except rose and this for some undetermined reason shows a minor decline. The average improvement for the eight odors combined is 21 per cent but the range is wide since it runs from—2 per cent (for Rose) to +53 per cent (for Pine).

When one considers the contrast between the number of correct selections made by each of the subjects before and after the identities of the floral odors were given, one finds that while the mean number of correct responses has advanced from 1.8 to 3.5, there is still considerable room for improvement. It is interesting to note that while none of the 69 subjects was able to identify more than four of the odors prior to their identities having been made known, 10 of the 35 subjects were able to do so after the list was made available. Two of the 35 were able to identify all eight samples correctly.

Summary and Conclusions

Employing 69 college students as the experimental group an attempt has been made to evaluate some of the factors that play a part in odor preferences and identifications. The results obtained are not intended to indicate universal trends, since a select group was used, but they do point to the need for further investigation in this area.

1. For the experimental group used the ability to recognize the difference between expensive and inexpensive perfume compounds was only slightly better than chance, with the mean percentage of correct responses being 55.

2. There was a greater tendency to select expensive perfumes as being inexpensive than vice versa.

3. Length of use of perfumes apparently does not affect the ability to make accurate judgments as to the costliness of perfume compounds.

4. Frequency of use does not affect the ability to make accurate judgments as to the costliness of perfume compounds.

5. There is considerable disagreement between the individual's evaluation of the cost of a perfume and its "pleasantness." There was a slightly greater tendency to attribute unpleasantness to odors thought to be costly than to consider as pleasant those compounds which were thought to be inexpensive.

6. Utilizing eight common floral odors it was found that our experimental group was able to identify them with less than 25 per cent accuracy (23.5 per cent correct).

7. When 35 subjects were informed as to what eight floral odors were being utilized their accuracy in identification rose to but 44 per cent.

Prediction of Female Readership of Magazine Articles *

Evelyn Perloff

Ohio State University

This is the second of two studies attempting to predict the number of individuals that will read a magazine article, prior to its publication. The first study discussed the prediction of male readership of articles in *The Saturday Evening Post.* The purpose of the current study was to determine the way in which five variables combined for maximum female readership of articles in the *Post.* The reader who desires complete details of the procedure used in these studies will find an account in the reference cited below.[1]

The readership results of men and women were handled separately on the assumption that interest patterns for magazine articles are well defined according to sex. A comparison of the readership figures of *Post* articles for males and females in these studies and many others will clearly illustrate the varying interests and preferences of the two sexes.[2]

Inasmuch as starting readership is based upon information obtained from individual reports from respondents, it is essential to have some measure of the accuracy of these reports. Ludeke and Inglis compared the results of what readers of the *Ladies' Home Journal* stated they had read with what they were observed to have read. The results of this informative experiment showed an average difference of 1.7% between the two conditions, which seems to justify the conclusion that "reported reading behavior did not differ materially from active reading behavior."[3] It is likely that similar results would be obtained with *The Saturday Evening Post.* At present, the reliability of the criterion probably lies within an error range of 8% (2σ value).[4]

Results

The five variables used were *number of illustrations, color of illustrations, sex of persons in the illustrations, proportion of opening page*(s) *devoted to text,* and *subject matter* of the article. The findings will be presented in three sections: (1) The Distributions, (2) The Determination of the Composite Effect, and (3) The Cross-validation.

The Distributions. All starting readership per cents are indexes and not actual figures.

The relationship ($r = .35$) of *number of illustrations* to starting readership per cent indicated on the face of it that the number of illustrations significantly influenced the female reader in starting an article. It was apparent that there were no clear-cut breaks in the distributions, as was present in the study on male readership. Although there was a slight upward trend in starting readership from articles having no illustrations to those having eight or more, this trend was not very distinct. It was clear, however, that female *Post* readers preferred articles with many illustrations as compared to those with no illustrations. Both men and women were equally influenced by this variable (r's = .35) when its effect on starting readership was determined, but all other variables were permitted to vary.

There appeared to be a clear-cut relationship ($r = .42$) between the *color of illustrations* and starting the article. There were two definite breaks for the four categories in this variable. Thus, there were sharp changes from "other" to the two categories, black and white and duotone; and from black and white and duotone to full-color. It was clear that the women in this study did not differentiate between black and white and duotone but keenly preferred articles having full-color illustrations. The *color of illustrations* seemed to be of greater importance to women ($r = .42$) than to men ($r = .28$) in influencing them to start reading articles in *The Saturday Evening Post.*

The relationship ($r = .38$) between *sex of persons in the illustrations* and starting readership also appeared to influence significantly the starting readership of *Post* articles by women readers. Apparently, the woman reader of the *Post* preferred any type of illustration other than that showing only men. The woman reader preferred illustrations including both males and females to illustrations including females alone, but this preference was slight. Again, female readers seemed to be more influenced by *sex of persons in illustrations* than male readers ($r = .22$).

There appeared to be a significant inverse relationship ($r = -.36$) between *proportion of opening page*(s) *devoted to text* and how many women would start to read an article. It was apparent that devoting

less than 20 per cent of the opening page(s) to text resulted in the highest starting readership. The differences among the three classes (i.e. classes in terms of *amount* of space devoted to text) were clear-cut and more distinct than in the male readership study. The general trend was for starting readership to improve as the per cent of text on the opening page(s) decreased.

There was greater variation among the classes of the *subject matter* variable than of any other. A number of the categories had too few cases to merit consideration as a separate class. This eliminated various classes which are part of the gamut of subjects upon which *Post* articles are written. These articles were classified under the category, "Other." It was found, however, after completion of the male readership study, that the category, "Other," could be further broken down into eight additional subject matter categories, making a total of 24 classes in the subject matter variable as compared to 16 categories in the previous study.

By this revision, the correlation coefficient between subject matter and starting readership per cent was raised from .46 to .60, both coefficients indicating clearly that the subject matter of an article considerably influenced the female reader to start it. It was apparent that the women (as well as the men) who read *The Saturday Evening Post* have definite likes and dislikes of *Post* topics. Although there was a steady increase in starting readership from topics least liked to those best liked, there were also several sharp changes grouping together both similar levels of preferences and similar kinds of subject matter. The general trend was for female starting readership to improve significantly when *Post* articles dealt with topics such as people at work, descriptions of peoples and places (USA), and health and hygiene. These topics revealed a preference by female readers for human-interest articles. Action-type articles such as those on sports, athletes, and labor, which rated among the highest with male *Post* readers, offered less attraction to the women readers.

The Determination of the Composite Effect. The correlation matrix is shown in Table 1. The horizontal and vertical headings indicate the five variables used in the study. *Number of illustrations* gave the lowest correlation ($r = .35$) with starting readership per cent, while the coefficient between subject matter and starting readership was the highest ($r = .60$).

Each of the five variables correlated higher with the criterion (starting readership) for female readers of the *Post* than for male readers. The writer is unable to say at this time whether this fact suggests that women,

Table 1

Intercorrelations Between Variables 1–5 and of Starting Readership Per Cent

(N = 190)

Variable	Starting Reader- ship %	No. of Illus.	Color of Illus.	Sex of Persons	% Text on Opening Page(s)	Subject Matter
Starting Readership %	—	.35	.42	.38	−.36	.60
No. of Illus.	.35	—	.67	.11	−.29	.19
Color of Illus.	.42	.67	—	.15	−.46	.33
Sex of Persons	.38	.11	.15	—	−.16	.25
Per Cent Text on Opening Page(s)	−.36	−.29	−.46	−.16	—	−.25
Subject Matter	.60	.19	.33	.25	−.25	—

by and large, are more impressionable than men, or more consistent in their interests, or were more greatly influenced by the particular variables used in the study, or whether this increased relationship (over male readership) resulted from the author's coding. It is probably safe, however, to conclude from the data in both this and the male study that there is a significant sex difference in the readership habits of *Post* articles.

For prediction purposes the regression equation was computed. Table 2 shows the weights that each variable obtained. These weights are an approximation of the relative independent value of each variable to the success of the article (starting readership per cent). Use of this regression equation yielded an R of .70. The standard error of estimate was 9.7 per cent. Hence, the chances are that in about 68 out of 100 cases the predicted starting readership per cents will be within an error of 10 points or less. We may be certain that very few starting readership estimates will be in error by more than 30 per cent.

Calculation of the coded score weights (weights dependent upon the

Table 2

Weights of Five Variables for Predicting Starting Readership Per Cent

(N = 190)   (R = .70)

Variables	Weight
Subject Matter	.45
Sex of Persons in Illustrations	.22
Number of Illustrations	.15
Proportion of Opening Page(s) Devoted to Text	−.14
Color of Illustrations	.07

measuring scale of the specific variable) gave the necessary data for the regression equation. The final equation is as follows:

Predicted Starting Readership Per Cent Index

= 14.9 (Index) + 1.9 × class value (No. of Illus.)

+ 2.3 × class value (Color of Illus.)

+ 6.1 × class value (Sex of Persons in Illus.) − 3.1 × class value (Proportion of Opening Page[s] Devoted to Text) + 4.2 × class value (Subject Matter).

Inasmuch as the correlation coefficients of four variables (not including subject matter) and the resulting multiple, .70, could be higher for predictive purposes, it is believed that there are other variables which possibly are of greater importance for prediction than the ones under present analysis. This is particularly evident from the fact that the multiple correlation coefficient was raised only .10 when these four variables were considered along with the subject matter variable. In view, however, of the paucity of information (i.e. other variables) and perhaps the difficulty of measuring them, consideration of the present variables, individually but more requisitely all together, can make noticeable improvement in predicting starting readership.

The Cross-validation. To determine the extent to which the weights of the characteristics of articles would be valuable in years other than the year 1946, when the articles included in this study appeared, we have applied this regression equation to 149 articles appearing in the 1947 issues of the *Post*. The correlation between the actual and predicted starting readership per cents was .73. This validity coefficient was slightly higher than the multiple ($R = .70$) and a reversal of the lower validity coefficient ($r = .36$) and the multiple ($R = .56$) obtained in the male readership study.

The higher correlation predictions for women readers in this later year probably result from a constancy in interests over the period of the year intervening. The increased number of classes in the subject matter variable may also account for the slightly higher validity correlation coefficient.

The average difference between the actual and predicted starting readership per cents was 8.6 per cent. The predicted starting readership per cents were within 5 per cent of the actual starting readership in 41 per cent of the articles, within 10 per cent in 68 per cent of the articles, and within 15 per cent in 86 per cent of the articles.

The Applications

The applications of this study are identical to those discussed in the study on male readership. The primary application lies in checking the

value of a tentative layout for an article and making such layout changes as are necessary to increase the average readership of each issue of the magazine. Inasmuch as weights may change with time, continued follow-up is essential.

Conclusions

The following conclusions are supported:

1. The multiple correlation and regression technique proved to be a successful method for predicting starting readership of *Post* articles by female readers.

2. The accuracy of the predictions of future articles should fall within a 10 per cent difference between predicted and actual starting readership per cents in about 68 per cent of the cases. This percentage error is satisfactory for most practical purposes.

3. The order of the relative importance of the five variables included in this study is (a) *subject matter;* (b) *sex of persons in illustrations;* (c) *number of illustrations;* (d) *proportion of opening page(s) devoted to text;* and (e) *color of illustrations.*

Special Review

Buros, Oscar Krisen. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. Pp. xv, 1047. $12.50.

Colossal is the word for this sixth and latest offering in the familiar series of bibliographical works on mental measurements edited by Buros. Starting with a modest 44-page listing of tests in 1935, the phenomenal growth of the series is charted in the following healthy chronology:

> 1935—Educational, Psychological, and Personality tests of 1933 and 1934—44 pages
>
> 1936—Educational, Psychological, and Personality Tests of 1933, 1934, and 1935—83 pages
>
> 1937—Educational, Psychological, and Personality Tests of 1936—141 pages
>
> 1938—The Nineteen Thirty-Eight Mental Measurements Yearbook—415 pages
>
> 1941—The Nineteen Forty Mental Measurements Yearbook—674 pages
>
> 1949—The Third Mental Measurements Yearbook—1047 pages

As indicated, the rate of increase in volume has been tremendous and as yet shows no clear change in trend. One is reminded of the fable of the sorcerer's apprentice; and one hopes that Buros has his wonderful editorial legerdemain under better control.

The *Third Yearbook* follows the familiar pattern of the earlier models. There are two main sections: "Tests and Reviews" and "Books and Reviews." The first of these, comprising over two-thirds of the volume, is the section for which the *Yearbooks* are best known, and on which their reputation is founded. The comparative statistics for the "Tests and Reviews" sections of the three *Yearbooks* are shown in the accompanying table.

Comparative Statistics of the Three Mental Measurements Yearbooks

	1938	1940	Third
Period Covered	Jan. 1937–June 1938	July 1938–Oct. 1940	Oct. 1940–Dec. 1947
Entries	313	503	705
No. of Reviews	331	503	713
No. of Reviewers	133	250	320

It was originally planned that new editions of the *Yearbook* would be issued at two-year intervals, but this schedule was interrupted by the war. The first post-war model, therefore, covers a period of some seven years—a fact which, in part at least, accounts for its gargantuan dimensions and excuses its minor sins of omission. The "Tests and Reviews" section of the *Third Yearbook* lists 663 tests (plus 42 references to books about single tests, e.g. the Rorschach) of which about 70 per cent are accompanied by one or more original reviews. Altogether, 713 reviews are contributed by 320 psychologists, educationists, subject-matter experts, classroom teachers, and test technicians. Included also are 66 excerpts from reviews which have already appeared elsewhere and (as claimed in the preface—this reviewer did not count them) "3,368 references on the construction, validity, use, and limitations of specific tests."

The tests listed and reviewed purport to be all of the "commercially available tests—educational, psychological, and vocational—published as separates in English-speaking countries between October 1940 and December 1947." In addition are included a selected list of "classics" (e.g. the Army Alpha, Stenquist Mechanical Aptitude, Strong Vocational Interest Inventory, etc.) plus a few tests published during the 15 years since Buros first started fishing in these waters, but which somehow "got away" before.

For each test entry, the following useful information is provided, all condensed into a few lines: test title; description of groups for which intended; date of publication, copyright, or revision; whether or not machine scorable; whether individual or group test; forms, parts, and levels available; cost; testing time and total administration time; author; and publisher. A specimen entry is:

> American Council on Education Psychological Examination for High School Students. Grades 9-12; 1933-1947; new form issued annually; IBM; separate answer sheets must be used; $2.00 per 25 tests; 50¢ per 25 machine-scorable answer sheets; 50¢ per specimen set; 35 (65) minutes; L. L. Thurstone and Thelma Gwinn Thurstone; Educational Testing Service.

Following this outline, for most test entries, are cross references to earlier *Yearbooks* or bibliographies in the series, and references to books and articles covering some aspect of the test. Then comes the feature which the series is best known—the original review. About a third of the entries include reviews by more than one contributor. But more about the reviews later.

The second main section of the volume—the "Books and Reviews" section—lists "549 books on measurements and closely related fields," accompanied in most cases by excerpts from reviews of these books culled from the journals. Here again the attempt has been to include

all works in the area published in English-speaking countries between October 1940 and December 1947. And here, as throughout the book, the emphasis is on critical evaluation. The editor's preface states in this connection:

> "Reviews which included no critical comment are listed but not excerpted. Readers should note that the critical portions of all book reviews, regardless of merit, found in professional and scholarly journals are included in this yearbook. Asterisks and ellipses within excerpts indicate the omission of non-evaluative material which appeared in the original review. . . ."

The selection of this material was presumably the sole responsibility of the editor. Inclusion of all the words written about all of the books listed would obviously have been both impossible and ridiculous. On the other hand, wherever judgment must be employed in the selection and editing of material, one is entitled to ask questions about the basis of the selection, and to be suspicious of possible conscious or unconscious bias. In this case, the editor assures us that *all* the *critical* appraisals of the books listed, collected from all the reviews of these books appearing in the professional and scholarly journals, have been included, and only the purely descriptive or non-evaluative reviews or parts of reviews have been left out. This being true, the reader might conclude that Buros' own *Nineteen Forty Mental Measurements Yearbook*, since it accounts for 50 reviews in 15½ pages, was either the most important or the most criticized—surely the most controversial—book on the subject published since 1941. Sharers of the honor of evoking the most critical comment would be the two volumes of *Diagnostic Psychological Testing* by Rapaport et al. (13 reviews, 16½ pages) and Stoddard's *The Meaning of Intelligence* (13 reviews, 10 pages).

In addition to the two major review sections covering tests and books, the volume contains five directories and indices. The first of these, *Periodical Directory and Index*, serves both as a key to the abbreviations used throughout in journal references and as a directory of journal editors. The second, *Publishers Directory and Index*, gives addresses of test and book publishers. The *Index of Titles* and the *Index of Names* are conventional alphabetical listings. Finally, the *Classified Index of Tests* is an expanded table of contents for the "Tests and Reviews" section, listing each entry numerically (1–705).

The reputation of the earlier *Yearbooks* derived principally from the "Tests and Reviews" sections; and the same will doubtless be true of the latest volume in the series. The major rubrics are essentially the same as those employed in the earlier issues; Achievement Batteries (22 entries); Character and Personality (91 entries); English (57 entries); Fine Arts (7 entries); Foreign Languages (36 entries); Intelligence (89

entries); Mathematics (62 entries); Miscellaneous, e.g. home economics,
safety, computational and scoring devices (111 entries); Reading (70
entries); Science (44 entries); Social Studies (30 entries); and Vocations
(86 entries). It is difficult to know what significance, if any, to attach to
the numbers of entries in each category. Perhaps they illustrate the
difficulties of fitting fabricated classifications to any given series of data,
particularly when the data present themselves without regard to the
principles on which the classification was originally compounded. When
this occurs, the pattern must be stretched here and there, and the miscel-
laneous section originally provided for overflow inevitably grows bigger
and bigger. In spite of this—and assuming the listings are as compre-
hensive as claimed—it appears that the war and post-war periods have
provided an atmosphere more congenial to production in the field of
character and personality than elsewhere. This conclusion might be
somewhat misleading, however, since one test alone accounts for nearly
a quarter of all the entries under this heading. Needless to say, it is the
Rorschach which somehow merits 67 pages, including a bibliography of
598 titles!

The original reviews themselves appear to this reviewer as a varied
lot having only one factor in common—all are critical. Criticism is, in
fact, the dominant tone of the whole volume (*nil nisi bonum* is definitely
not the editor's watchword!) and while going through it page by page
one may conjure up an image of the editor at the head of his ranks of
contributors daring the would-be test maker to attempt to get away
with anything shoddy or unscrupulous. The image is an inspiring one,
and, though fanciful, not too remote from the editor's intention. One
of the major objectives of the *Yearbooks*, in fact, is:

> "To impel authors and publishers to place fewer but better tests on the
> market and to provide test users with detailed and accurate information
> on the construction, validation, uses, and limitations of their tests at the
> time that they are first placed on the market."

To achieve this objective, the editor has instructed his cooperating
reviewers to provide reviews that are ". . . frankly critical with both
strengths and weaknesses pointed out in a judicious manner." Just
how "judicious" are such randomly selected remarks as:

> "This is just another test for neurotic tendencies. The reviewer can
> think of no reason for its publication or use. . . . The only excuse for
> publishing another test of neurotic tendency in this day and age is in-
> creased validity over other tests in the field. This test is grossly lacking
> in this respect.
>
> With the perspective attained in the years since its publication . . . one
> may view (test maker's) arrant nonsense with tolerant amusement."

or, again:

> "The instrument is a reversion to a type of psychological and sensory testing that belongs to the infancy of mental measurement, and has repeatedly been proved worthless as an index to higher mental ability."

the reader must judge for himself. If such candid criticisms are indeed warranted, the reviewers deserve full praise for their courage is saying so. Fortunately, a good many of the reviewers have displayed this kind of forthright frankness. Certainly, this reviewer is not advocating libelous brutality for the sheer sadistic enjoyment of contemplating the discomfiture brought about by a well-placed literary needle. But it is difficult to discern what value might accrue to the potential test user from that type of review which finds a little to praise and a little to blame in every test and sums up with an equivocal statement of possible usefulness as a crutch for intuitive hunches. Fortunaltely, this sort of review is not found very often in the *Yearbook*.

But while the general tone of the book is healthily evaluative, the basis of criticism varies considerably among the reviews. They might well be classified under headings suggested in an excellent review of the *1940 Yearbook* (Pedro T. Orata in *The Teachers College* Journal (Manila), 1941, 3, 59–61):

1. Emphasizes functional or "true" validity, criticizes test for success or failure to measure ultimate educational objectives . . . in measuring worthwhile results of instruction; and in general subordinates the techniques of test construction and mechanics of administration, scoring, and tabulation of scores to the higher values that the test and testing should engender in the pupils and in those who use it.
2. . . . criticizes the test mainly for success or failure to meet the requirements of statistical validity and reliability, evaluates it on the basis of commonly accepted techniques of test construction, and in general assumes functional validity or subordinates it to formal content and make-up of the test.
3. . . . evaluates it mainly from the point of view of its success or failure to meet the mechanical requirements of efficiency in scoring, administration, and tabulation of test results.

Each of these three points of view has merit, to be sure, and each could doubtless rally a considerable number of supporters to its side. In fact, it is probably more to the point to classify reviewers in this manner than to classify their products. And herein lies the fundamental weakness of the *Yearbook* in the opinion of this reviewer. Granting that all of the contributors to the volume are experts and qualified to speak and be heard, they are not all equally sensitized to all phases of psychometry. In diagnosing human ailments, we don't call in only the internist or the neurologist. Nor should we, in examining the test, expect the subject-matter expert to detect statistical ailments, or the psychometrist to point up an undernourished teaching objective. We should call in all the

specialists and hold a thorough clinical examination on each case. That something like this was intended is indicated by the editor's statement of objectives in the *1940 Yearbook:* to provide reviews "written by persons of outstanding ability representing various viewpoints. . . ." But in calling in the experts, the editor has made the assignments of cases, which implies that he already knew the patients' needs. Though some such procedure is a practical necessity in a venture of this kind, it has the definite disadvantage that the treatment of the various types of tests is apt to be unbalanced. A cursory survey, for example, indicates that nearly all the reviewers assigned to the Achievement Batteries are educationists, educational researchers or examiners, and that psychologists and psychometrists predominate among the reviewers of Intelligence tests.

After many hours spent in contemplating the somewhat frightening aspect of the *Third Yearbook*, this reviewer found himself musing about the practicability of another kind of volume. This "dream" book would not attempt to reproduce verbatim the literary efforts of the experts, but would edit and cull from them all the essential materials to fill out a standard outline. Spared the necessity of literary composition, reviewers could concentrate on specifics, and could handle more tests with no greater expenditure of effort. The outline itself would be drawn up by a board of outstanding specialists including both test makers and test users. It would cover such points as: type of item; sources of items; nature of item analysis; descriptions of populations used for item analysis, factorial analysis, validation, cross-validation, standardization; judgments of functional validity; adequacy of "coverage"; et cetera, et cetera. This list is obviously not exhaustive, and many readers may detect a statistical bias in it. It is for precisely this reason that the board of experts would be used to insure the inclusion of all important dimensions of a test. Finally, there would be the main feature of the book—the board of experts' "seal of approval" for tests which merited adoption and use. In this last connection, the reviewer is reminded of the statement made by Sandiford in commenting on the earlier *Educational, Psychological, and Personality Tests of 1936* (American Journal of Psychology, 1938, 51, 200): ". . . Professor Buros' annual publication would be made much more useful if he would mark with a prominent star those (tests) which were valid, reliable, and had satisfactory norms. Then busy workers could neglect the rest, or if they wasted their money on 'gold bricks,' the fault would be their own." This reviewer can think of no better way of achieving the objectives of fewer and better tests.

E. Donald Sisson

Personnel Research Section, AGO,
Department of the Army.

Book Reviews

Ahern, Eileen. *Survey of personnel practices in unionized offices.* Research report number 13. New York: American Management Association. 1948. Pp 38. $1.50 (non-members, $3.00).

This report consists of twenty frequency tables and accompanying text relating to practices in unionized offices in matters of union security, salaries, hours of work, leave of absence, group insurance, seniority, discharge, grievance adjustment, and other collective bargaining subjects. The report is based on 50 union contracts believed to be fairly representative of the entire AMA collection of 300 office union contracts.

The report will be of interest to only a few psychologists. Those who are concerned with collective bargaining with office unions or those who wish to compare their practices with those obtained by employees through collective bargaining will find the report of some interest subject to the limitations imposed by a sample of 50 cases and sub-group tabulations based on an N ranging from five to eight.

<div style="text-align: right">C. E. Jurgensen</div>

Minneapolis Gas Company

Achilles, Paul S. *Management and the psychologist: A practical guide on psychology for the business executive.* Section II, Book 4, Reading Course in Executive Technique, Ed. by Carl Heyel. New York: Funk and Wagnalls Co., 1948. Pp. 64. $1.00.

The sub-title is an exact description of the contents of this little book. The presentation is concise yet it is surprisingly comprehensive. It is authoritative, readable and accomplishes its purpose in admirable fashion. It is just the type of book to place in the hands of the business executive who has never been exposed to formal psychology but who may be curious as to just what our discipline is all about.

Only one minor criticism would appear to be justified. Having whetted the appetite of the business executive, Achilles might well have added a short selected annotated bibliography for his guidance in case he might desire to pursue further any phase of the subject.

<div style="text-align: right">Donald G. Paterson</div>

The University of Minnesota

Linebarger, Paul M. A. *Psychological Warfare.* Washington, Infantry Journal, 1948. Pp. 259. $3.50.

The purpose of this book is to tell a layman audience what psychological warfare is and how it is fought. Linebarger is Professor of Asiatic

Politics at the Graduate School of Advanced International Studies in Washington, D. C. He served in the War Department and in OWI in both policy formulation and in field operations.

The book handles psychological warfare in three parts. Each part has three to six chapters. In the first part, Linebarger covers historical examples, definitions, limitations and characterizations of national uses of psychological warfare in World Wars I and II. The second part is devoted to how to analyze and derive military intelligence from propaganda in order to make an objective appraisal of a given situation in terms of psychological warfare. The third phase includes organization, plans, operations, and remarks on future problems.

The strong point of the book is the waggish style. This is exemplified when he pokes fun at the high level policy echelons wherein much of the output was classified top-secret and thus removed from usefulness. There are seventy excellent figures of propaganda leaflets as well as ten organizational charts of various national offices involved in psychological warfare. The content is enlivened with descriptions of events such as the use of radio-phones in tank warfare to induce Japanese surrenders. The three major U. S. lessons from World War II are, he says, that atrocity propaganda does not pay, that we have no backlog of trained propaganda personnel, and that psychological warfare must be a positive function at command level, not a sideline specialty apart from top level policy making.

A weak point of this book is its lack of organization despite the promise of the excellent chapter headings. Specific techniques, the root of the entire matter from a professional view, are mentioned as the story develops. They are not consolidated for a comparative analysis of their uses and limitations. There is an unusual mixture of the levels of vocabulary. Such words as condign, maleficent, and oestrous occur as well as frequent references to people going mad with confusion and serious use of Frisco for San Francisco. Use of the revised Flesch formulas show readability as difficult and style as mildly interesting. Linebarger epitomizes and tends to rest content with neatly turned phrases. For example, he makes the point that education is to psychological warfare what a glacier is to an avalanche. He neglects to show the crucial differences in bias, in use of segmental appeals, and in emotional and authoritarian contexts. Professional psychologists may wonder if his two page discussion of the role of the psychologist in warfare justifies use of "psychological" in the title. The location of the eighty illustrations with reference to the text might have been improved.

In summary, Linebarger's book presents "a patchwork of enthusiastic recollection" as he calls it. Although some professional readers

may be disappointed, the fact that it is a lively entry in a relatively undeveloped field makes the book worthwhile for his intended audience.

Clark L. Hosmer

Lt. Col. U. S. Air Force

Terman, L. M., and Oden, Melita H. *The gifted child grows up: Twenty-five years' follow-up of a superior group.* Stanford, California: Stanford Univ. Press, 1947. Pp. xiv, 448. $6.00.

As stated in the preface, the volume "is an over-all report of the work done with the California group of gifted subjects from 1921 to 1946, the greater part of it being devoted to a summary of the follow-up data obtained in 1940 and 1945; at the latter date the average age of the group was approximately thirty-five years."

The first six chapters are a resume of the earlier work, reported in more detail in two previous monographs. When selected in 1921–3, the 1050 pre-high-school subjects had an average chronological age of 9.7 years, and the 420 high school cases, 15.2 years; I. Q.'s ranged from 135 to 200 with a mean of about 150. It was estimated that the group was in the highest one per cent in ability as measured. Thirty-one per cent of the fathers were professional men; 60% of the homes were rated as superior; relatives included many individuals of note. In 1923, thirty-seven anthropometric measurements of 59 per cent of the cases showed that "in all respects . . . the selected group was slightly superior physically to the various groups used for comparison." Health histories and medical examinations showed health to be better and defects less common as compared with the average child; puberty tended to be reached a little earlier. In school 85 per cent were accelerated in grade placement; nevertheless, tests showed over half to have mastery of subject matter two grades yet further ahead. Interests of the gifted were livelier, more mature, more intellectual, and somewhat more social than for average children. Tests and ratings of character traits also showed superiority of the gifted. A second survey six years later yielded results substantially in agreement with the first findings.

Chapters 7–19 are concerned with follow-ups in 1940 by inquiry forms and field workers where possible, and by inquiry forms in 1945–6. That cooperation was outstanding is indicated by returns of information from 93 per cent of all living subjects. Mortality was to date found less than for the general population, physique and health superior, and maladjustment, delinquency and insanity less frequent than in the general population. Of the total group, 70 per cent of the men and 67 per cent of the women had by 1945 graduated from college (as compared with 5 per cent of the general population); 34 per cent of the men took one or more

graduate degrees; academic records were superior, median age of graduation was over a year younger than usual. Nevertheless, gifted students participated more in extra-curricular activities than the average student. Approximately 71 per cent of the men were in professional or superior business occupations in 1940, or 5 times as many as for California men in general; income was higher than for college graduates in general. Avocational interests were diverse and rich. Attitudes were middle-of-the-road. More had married and at earlier ages than for college graduates in general, but divorce was less than half as frequent; happiness in marriage was rated high, and sex adjustment appeared in no way atypical.

Chapters 20–26 deal with somewhat special problems. Accelerates in school were found greatly to excel non-accelerates in the group, in achievement on a test battery in 1922; in 1940, over twice as many accelerates were in the top group in vocational success. Accelerates married earlier, and appeared not handicapped in adjustment or in physical or mental health. Special study of the subjects with I. Q.'s of 170 or above show them "about as successful as lower testing subjects in social adjustments," and they accomplish more. Subjects of Jewish descent differed little from the non-Jewish "except in their greater drive for vocational success, their somewhat greater tendency toward liberalism in political attitudes, their somewhat lower divorce rate." A vigorous chapter on factors in the achievement of gifted men showed the most successful distinguished primarily not by intelligence but especially by drive to achieve, and by all-round adjustment; outstanding accomplishment was not associated with marked emotional tensions but rather with stability and freedom from excessive frustration. War records were good. A careful chapter on the appraisal of achievement emphasized the variety of possible values, the possibility that admirable achievement might not involve eminence, and the need for later data if appraisals of accomplishment are to be adequate. The final chapter stresses the importance of future follow-ups, and over-views the total investigation in larger perspectives.

In total, then, the volume is an outstanding example of that most rare, but probably most valuable type of psychological investigation—the broadly conceived, long-time developmental study. The subjects were that portion of the total population most valuable to society. For all who are interested in problems of human personality in its finest potentialities, or the most challenging opportunities in education and guidance, the volume should be a "must."

<div style="text-align: right">Sidney L. Pressey</div>

Ohio State University

New Books, Monographs, and Pamphlets

ABC's of scapegoating. Revised edition. Gordon W. Allport. New York: Anti-Defamation League of B'Nai B'rith, 1948. Pp. 56. $.20.

Practical psychology. Karl S. Bernhardt. New York: McGraw-Hill Book Co., Inc., 1948. Pp. 319. $2.50.

Industrial psychology and its social foundations. Milton L. Blum. New York: Harper and Brothers, 1949. Pp. 518. $4.50.

Student personnel services in general education. Paul J. Brouwer. Washington, D. C.: American Council on Education, 1949. Pp. 317. $3.50.

Managers, men and morale. Wilfred B. D. Brown and Winifred Raphael. London, Eng.: MacDonald and Evans, 1948. Pp. 163. 10/6.

The third mental measurements yearbook. Oscar K. Buros, Editor. New Brunswick: Rutgers University Press, 1949. Pp. 1047. $12.50.

Personal adjustment in old age. Ruth Shonle Cavan, Robert J. Havighurst, Ernest W. Burgess, and Herbert Goldhamer. Chicago: Science Research Associates, 1949. Pp. 175. $2.95.

The psychology of social classes. Richard Centers. Princeton: Princeton University Press, 1948. Pp. 432. $5.00.

Psychologist unretired. Miriam Allen deFord. Stanford: Stanford University Press, 1948. Pp. 127. $3.00.

Film and education. Godfrey Elliott, Editor. New York: Philosophical Library, 1948. Pp. 597. $7.50.

The energetics of human behavior. G. L. Freeman. Ithaca: Cornell University Press, 1948. Pp. 352. $3.50

Man's place in God's world. Sol W. Ginsburg. New York: Hebrew Union College, Jewish Institute of Religion, 1949. Pp. 30. $.50.

4-Square planning for your career. S. A. Hamrin. Chicago: Science Research Associates, 1948. Pp. 200. $2.95.

Adolescent character and personality. Robert J. Havighurst and Hilda Taba. New York: John Wiley and Sons, Inc., 1949. Pp. 315. $4.00.

How to create job enthusiasm. Carl Heyel. New York: McGraw-Hill Book Co., Inc., 1948. Pp. 248. $3.00.

Psychology and ethics. Harry L. Hollingworth. New York: Ronald Press Co., 1949. Pp. 247. $3.50.

Applied psychology. Revised edition. Richard Wellington Husband. New York: Harper and Brothers, 1949. Pp. 845. $4.50.

Theory and problems of social psychology. David Krech and Richard Crutchfield. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 639. $4.50.

Discovering your real interests. G. Frederic Kuder and Blanche B. Paulson. Chicago: Science Research Associates, 1949. Pp. 48. Single copy, $.75. Fifteen or more copies, $.60.

Personality projection in the drawing of the human figure. Karen Machover. Springfield, Ill.: Charles C. Thomas, Publisher, 1949. Pp. 181. $3.50.

Psychological statistics. Quinn McNemar. John Wiley and Sons, Inc., 1949. Pp. 364. $4.50.

Workers wanted. E. William Noland and E. Wight Bakke. New York: Harper and Brothers, 1949. Pp. 224. $3.00.

The procurement and training of ground combat troops. Robert Palmer, Bell I. Wiley, and William R. Keast. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1948. Pp. 696. $4.50.

Industrial hygiene and toxicology. Volume I. Frank A. Patty. New York: Interscience Publishers, Inc., 1948. Pp. 531. $10.00.

Machine computation of elementary statistics. Katharine Pease. New York: Chartwell House, Inc., 1949. Pp. 238. $2.75.

Job horizons. Lloyd G. Reynolds and Joseph Shister. New York: Harper and Brothers, 1949. Pp. 102. $2.25.

Human relations in an expanding company. Frederick L. W. Richardson and Charles R. Walker. New Haven, Connecticut: Yale Labor and Management Center, 1948. Pp. 95. $1.50.

Company annual reports to stockholders, employees, and the public. Thomas H. Sanders. Boston: Division of Research, Harvard Business School, 1949. Pp. 338. $3.75.

An outline of social psychology. Muzafer Sherif. New York: Harper and Brothers, 1948. Pp. 479. $4.00.

Government regulation of industrial relations. George W. Taylor. New York: Prentice-Hall, Inc., 1948. Pp. 383. $4.00.

Social class in America. W. Lloyd Warner and Kenneth W. Eells. Chicago: Science Research Associates, 1949. Pp. 292. $4.25.

Constructing classroom examinations. Ellis Weitzman and Walter J. McNamara. Chicago: Science Research Associates, 1949. Pp. 140. $2.50.

Human behavior and the principle of least effort. George Kingsley Zipf. Cambridge: Addison-Wesley Press, Inc., 1949. Pp. 650. $6.50.

Symposium on industrial relations. American Journal of Sociology. January 1949 issue. Chicago: University of Chicago Press, $1.25.

Factors affecting the satisfactions of home economics teachers. AVA Research Bulletin No. 3. Washington, D. C.: Committee on Research and Publications, American Vocational Association, Inc., 1948. Pp. 96. $.75.

Hours of work and output. Bulletin No. 917. Bureau of Labor Statistics. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1948. Pp. 160. $.35.

Operating under the LMEA, relation of wages to productivity. Personnel Series Number 122. New York: American Management Association, 1948. Pp. 63. $1.25.

Sociometry and group relations. Work of Progress Series. New York: American Council on Education, 1948. $1.25.

The open house in industry. Chicago: National Metal Trades Association, 122 South Michigan Avenue, 1948. Pp. 27.

The UAW-CIO looks at time study. Detroit: UAW-CIO Education Department, 28 West Warren Street, 1947. Pp. 32. $.50.

Employees suggestion programs in the iron and steel industry. New York: American Iron and Steel Institute, 350 Fifth Avenue, 1948. Pp. 92.

Air conditioning in textile mills: the case for temperature and humidity control to provide comfort, health, safety, and optimum production. New York: Research Department, Textile Workers Union of America, 99 University Place, 1948. Pp. 60.

A Re-examination of the Accident Proneness Concept

Alexander Mintz and Milton L. Blum

College of the City of New York

It is generally accepted that certain individuals consistently have many accidents while others do not. This is commonly known as the principle of accident proneness. A critical examination of the data reported in the literature points to the desirability of reconsidering the significance attached to the principle of accident proneness.

This article has two objectives: (1) To indicate that one of the methods to substantiate the principle of accident proneness is unsound and to show that its use has led in some instances to exaggerated views of differences in accident proneness; and (2) To propose a method whereby quantitative estimates of differences in accident liability [1] may be obtained and to point out the conditions when it may be used.

The statistical evidence for the principle of accident proneness was presented by Greenwood and Woods (6) in 1919. These authors compared the distribution of accidents in a given population with a simple chance distribution for the same number of accidents in a population of the same size. Evidence of differences in accident proneness was obtained: It was discovered that more people had no accidents than might have been expected "by chance." Conversely, it was discovered that more people had many accidents than would have been expected in accordance with a simple chance distribution. In other words, Greenwood and Woods demonstrated that the obtained distributions of accidents differed significantly from chance expectancy. Furthermore, they showed that most of their distributions agreed with theoretically computed distributions based on the assumption that people differed from each other in their likelihood to have accidents.

Newbold (9) further investigated this problem and pointed out that the differences in accident liabilities could not be entirely explained simply

---

[1] In the subsequent discussions we shall use the expression "accident proneness" in referring to personal characteristics of people contributing to the likelihood of their having accidents. The expression "accident liability" will refer to both personal characteristics and stable environmental conditions contributing to accidents records.

in terms of different job hazards. In addition, Newbold, in some of her work, compared the accident rates for the same people in two successive periods and reported that significant correlations existed.

Both Greenwood and Woods, and Newbold were primarily interested in the establishment of the existence of a difference between accident records and chance expectancy. In this they were successful and accordingly the principle of accident proneness was established.

However, another method has been used to support the principle of accident proneness. A number of investigators and writers of books on industrial psychology have pointed out that small percentages of people have large percentages of accidents and have presented data accordingly. In this method the obtained accident distribution is presented as evidence for the principle of accident proneness without a comparison to the distribution that would be normally expected "by chance," i.e., if all individuals were equally liable to accidents. This method is fallacious.

The Method of Percentages

The method of percentages of people and accidents implies an incorrect assumption, viz., that chance expectation requires that all people in a population should have the same number of accidents. This is not the case. An obvious limitation that has often been overlooked is the fact that very often the reported total number of accidents in a population is smaller than the number of people in the population. For example, if a group of one hundred factory workers had fifty accidents in one year, then a maximum of fifty people could have contributed to the accident record and accordingly a maximum of 50% of the population would have contributed to 100% of the accidents. Obviously a small percentage of the population in this case does not establish the principle of accident proneness. However, the number of employees having accidents is almost certain to be less than fifty since there is no reason to believe that each one should have had only one accident. Such an assumption would imply that an accident immunizes its victim against further accidents. If one makes the assumption of equal liability, the people who had one accident should be just as liable to have future accidents as those who have not had any. Thus if accident liability is unchanged by accidents already had, some people should have two accidents before others have had any. In fact, in accordance with chance expectancy some people should have had three or more accidents before another had a single accident. In dealing a deck of cards it is not improbable that a person will receive more or less than the three or four cards in a suit that seem to be his share. He may get six, seven or more such cards without any laws of probability being violated. Similarly, a person

may have more accidents than seems to be his share in a given population without being more accident prone than the average.

Thus the assumption of equal accident liability results in different accident totals for the individuals within the group. The resulting distribution can be readily derived from the statement, "the current rate at which accidents occur per person is identical in groups of people with different numbers of accidents in the past." It follows directly from this statement that as the number of people who have had no accidents decreases, fewer people are likely to have first accidents per unit of time; as the number of people who have had first accidents increases, the rate of occurrence of second accidents increases proportionately. These and other similar statements can be reformulated as a set of differential equations, and the solution of this set of equations gives the terms of the Poisson distribution. Greenwood and Yule (7) first demonstrated its applicability to the accident problem. The Poisson is a discrete distribution rather than a continuous one. As applied to the accident problem, its consecutive terms give the predicted numbers of people who had no accidents, one accident, two accidents, etc. The terms are

$$Ne^{-m}, \ Ne^{-m}m, \ N^-e^m \frac{m^2}{2!}, \ Ne^{-m}\frac{m^3}{3!}, \ \text{etc., where } N \text{ is the number of people,}$$

e is the constant $2.71828\cdots$, m is the mean number of accidents per person.

A number of sets of data will now be discussed in order to illustrate the inadequacy of the method of percentages of people and accidents.

Based upon original records obtained by the authors from a foundry it was found that 1.8% of the 280 men in the day shift had 11.4% of the accidents; 10% of the men had 44.3% of the accidents. In the night shift 5.8% of the 120 men had 12.5% of the accidents and 37.5% of the men had all of the accidents. A computation of the distribution of accidents in accordance with chance expectancy (equal liability distribution) indicated that the differences between the obtained and expected distributions were not significant. In accordance with the theoretical distribution, 1.4% of the people should have had 8.3% of the accidents and 8.9% of the people should have had 38.8% of the accidents. These percentages obtained from a theoretically computed equal liability distribution show that the accident distribution actually obtained is in accordance with chance expectancy and does not establish the existence of accident proneness.

A study that is often referred to in discussions of accident proneness is that of the National Association of Taxicab Owners and the Metropolitan Life Insurance Company (11). These data deal with the records of 1294 drivers employed by several taxicab companies. Viteles (13)

states that "the incidence of accident proneness in the operation of motor vehicles has been well demonstrated in this study." "It is interesting to note that the data obtained in accident prone studies in other types of industries if plotted would closely conform to the curve shown. . . ."

Neither the authors of the report nor the author of the textbook compared the data with the simple chance distribution. Such computations have been made and are presented in Figure 1.
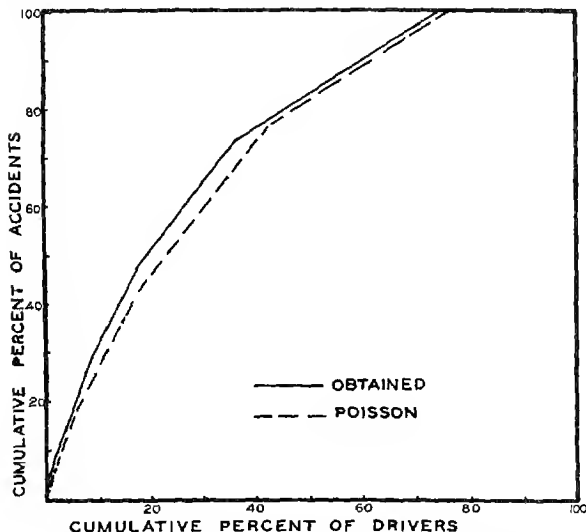
FIG. 1. Relationship between cumulative percentage of taxi drivers and of accidents.

The solid line in the figure represents the cumulative percentages of accidents corresponding to cumulative percentages of drivers, based on the data as quoted in the original report. The dotted line represents the corresponding cumulative percentages from an equal liability distribution.

The two lines are obviously very similar in shape. The argument (13) could be repeated verbatim with percentages from the chance distribution

substituted for obtained percentages, with very little loss in apparent persuasiveness. In the chance distribution, 23.5% of the people would have had no accidents instead of the obtained 25.2%. The best and the worst 50% would have had 18.3% and 81.7% of the accidents respectively, instead of the actually obtained 17.2% and 82.8%. The worst third of the drivers would have had 63.9% of the accidents (instead of 69.3%); the worst 10% would have had 24.7% instead of 31.9% of the accidents.

In spite of the fact that the two distributions are very similar in shape, the difference between them is statistically significant, the chi square being 122.77 (d.f. = 6, P < .0001). In other words, factors other than so called chance factors are definitely present but do not markedly change the general shape of the chance distribution.

Another often referred to study on accident proneness is the one reported by Slocombe and Brakeman (12). Their data are based upon accident records of 2300 men employed by the Boston Elevated Railway Company.

In discussing their data as indicative of differences in accident proneness, Slocombe and Brakeman classified the men with four or less accidents as "low accident men" and those with five or more accidents as "high accident men." This arbitrary division placed 1828 men in the first category and 472 men in the latter division. The "low accident" men averaged 2.1 accidents while the "high accident" men averaged 7 accidents. Slocombe and Brakeman did not compute the chance expectancy of the number of men having four accidents or less. Actually, in a simple chance distribution, 1824 men should be expected to be in this category and so only four more men of the total 2300 are in the "low accident" group than obtained by chance. According to chance expectancy, the "low accident" and "high accident" men should have averaged 2.4 accidents per man and 5.8 accidents per man respectively. The difference is not much smaller than the one actually obtained. This does not mean that there is no evidence for differences in accident proneness in the data. It merely means that Slocombe and Brakeman's line of argumentation is inconclusive.

More recent data based upon a random sample of licensed drivers in the state of Connecticut (2) have been analyzed by Cobb (1). He computed the amount by which the variance of accident records exceeds the variance of the Poisson distribution and thus determined that these accidents records cannot correlate with a perfect test of accident proneness to a degree higher than +.44.

DeSilva (2) refers to these data and uses as argument for the principle of accident proneness mainly the fact that four per cent of the drivers

were responsible for 36% of the accidents.   In a simple chance distribution 2.4% of the drivers would be responsible for 21.2% of the accidents. Again a comparison of percentages of people and of accidents in inconclusive.   The figures just quoted based on the assumption of a simple chance distribution look almost as impressive as the figures in the actually obtained distribution.[2]

Quantitative Estimate of Differences in Accident Liability

It is possible to arrive at an estimate of the magnitude of differences in accident liability (as distinguished from differences in accident records) in the case of many populations.   The procedure has been previously used by Cobb (1) as a step in estimating the maximum correlation between accident records and psychological tests.   This procedure can be used in many instances to estimate the magnitude of differences in accident liability, but it is also necessary to mention that this procedure is not universally applicable.

The presence of differences in accident liability of individuals in a population results in a composite of Poisson distributions of the accident records.   The reason for this is as follows: Each particular degree of accident liability present in a population should result in a Poisson distribution of the accident records.   When two or more degrees of accident liability are present the resulting distribution is the sum of the two or more corresponding Poisson distributions.   If the distribution of accident liability is a continuous function the resulting probability function of accidents is a composite of Poisson distributions which can be determined by integration.

When a given distribution of accident records is found to conform closely to a composite of Poisson distributions the evidence is consistent with the assumption that the differences between the accident records of different people are due partly to differences in their accident liability and partly to "chance" factors not predictable in terms of knowledge of the people or of their accident records.   In this assumption, the "chance" factors produce the variability within the constituent Poisson distributions while the differences in accident liability are responsible for the differences between their means.   In accordance with such an assumption, one may analyze the obtained variance of a set of accident records

into two constituent variances and view one of them as representing the operation of the "chance" factors, the other as characterizing the differences in accident liability. The former is the weighted arithmetic average of the variances of the Poisson distributions. As Cobb has shown, its value can be readily estimated as equal to the mean number of accidents per person.[3] Thus the residual variance representing the operation of differences in accident liability may be estimated if one subtracts the mean number of accidents per person from the obtained variance of accident records. We have performed this computation for a considerable number of accident distributions and have expressed the resulting variances attributable to unequal accident liabilities as percentages of the corresponding total variances of accident records.

The argument of the last paragraph pre-supposes that the obtained accident distribution approximates a composite Poisson distribution. Theoretically, an infinite variety of such distributions could be computed, depending on the assumed form of the distribution of the means of the Poisson distributions. Actually only one kind of such composites seems to have been used in research, viz., Greenwood and Yule's (7) "unequal liability distribution" ("UD"). This distribution is based on the assumption that accident liability of people is distributed along a Pearson Type III curve, a continuous skewed unimodal curve. Its equation may be found in several sources, e.g. (3), (8). Many sets of accident data can be actually approximated by composite Poisson distributions based on such assumed distributions of accident liability. It should be noted however, that Greenwood and Yule's "UD" distribution is by no means the only possible unequal liability (composite Poisson) distribution. Greenwood and Yule (7) report a set of equations for a different type of composite Poisson distribution, based on the assumption that accident liability is normally distributed. This distribution does not seem to have been used in research. The possibility should not be overlooked that this distribution or still another composite Poisson distribution, based on some other assumed distribution of accident liability, might prove to be useful in research. In this paper, composite Poisson distributions based on the Pearson III curve were used most of the time. In a few instances another possibility was explored to some extent; some sets of data suggested discontinuous distributions of accident liability, the discontinuity being due to the presence of small numbers of deviant individuals. On the other hand, the presented analysis of the sample variance into two components is not legitimate if the obtained distribution deviates significantly from any composite Poisson distribution.

[3] This follows from the fact that in a simple Poisson distribution the variance is always equal to the mean. Hence, in a composite of such distributions, the mean of the variances is equal to the mean of the means.

The line of reasoning just developed will now be applied to the more widely known studies of accident proneness.

The Greenwood and Woods study (6) presents fourteen sets of data. The majority of their findings agree rather well with the composite Poisson distributions computed according to Greenwood and Yule (7). In other words, the obtained figures are in accord with the assumptions:

1. Accident proneness varies from person to person and its distribution is represented by a unimodal continuous skewed curve known as Pearson type III.

2. Accident proneness of a person is unaltered by accidents he may have.

Twelve of the fourteen sets of data do not differ significantly from the corresponding theoretically computed figures. The P's reported by Greenwood and Woods obtained from the chi square technique range from 0.15 to 0.93.[4]   The sum of the chi squares for those 12 sets of data, based upon our computations, is 35.33, which for 30 degrees of freedom results in a P equal to about .23.   The two deviant distributions will be discussed later.

Thus it is possible to approximate closely the majority of Greenwood and Woods' tables by theoretically computed distributions based on the assumptions that accident proneness is constant for each person and distributed in different people in accordance with a Pearson III curve. This finding is one of the principal ones in favor of the existence of differences in accident proneness.

How large then are these differences in accident proneness if we take the findings at their face value and assume that variations in "chance" and differences in accident proneness are the only factors accounting for these distributions of accident records.   Table 5 presents .the data pertaining to the relative size of these differences in the Greenwood and woods study.

For each one of Greenwood and Woods' tables the estimated percentage of the variance of accident records attributable to differences in accident liability is given.   As stated on a preceding page, the estimated variance of accident liability is the difference between the obtained variance of accident records and the mean number of accidents.   Dividing this difference by the variance of accident records we obtain the percentage of the variance attributable to differences in accident liability.   In addition, the following data are also given: the number of cases, the mean and the variance of accident records.

---

[4] The computations do not appear to be accurate in all cases.   It is to be noted that the paper appeared in 1919 prior to Fisher's pointing out the procedure for determining degrees of freedom.

Table 5

Percentages of Variance Attributable to Differences in Accident Liability, from Greenwood and Woods Original Data

Greenwood and Woods Table No.	Number of Cases	Mean (m)	Obtained Variance (m_2')	$\dfrac{m_2' - m}{m_2'} \times 100$
I (A)	750	0.576	0.540	—
I (B)	580	0.478	0.491	—
II (A)	647	0.465	0.691	32.7%
II (B)	584	0.433	0.521	16.9%
III	100	3.040	6.938	56.2%
IV	414	0.483	1.008	52.1%
V	201	0.473	0.508	7.0%
VI	198	1.318	1.873	29.6%
VII (A)	59	0.983	1.203	19.3%
VII (B)	136	0.794	0.928	14.4%
VIII (A)	50	2.800	6.720	58.3%
VIII (B)	50	1.920	3.313	42.1%
IX (A)	55	2.473	3.704	33.2%
IX (B)	61	0.705	0.897	21.4%

The median percentage of the total variance, attributable to differences in accident liability is 31.15. The percentages range from 7% to 58.3%. In nine of the twelve cases the percentage is less than 50. These figures hardly correspond to the impressions one is likely to derive from textbook accounts. The share of differences in accident liability in the variance of accident records is very variable; it exceeds 30% in only half of the cases while the rest of the variance which is more than twice as large must be attributed to unpredicatble "chance" factors.

Newbold (9) collected a large number of sets of data from a number of factories. The factories were chosen on the basis of uniformity of the work performed, completeness of accident recording and opportunities for many minor accidents. The large majority of the accidents were trivial in nature, the author stating that the serious injuries were too few for correlational work. The findings differ in some respects from those of Greenwood and Woods.

A large variety of results can be found in Newbold's material. Nevertheless, in general the ratio $\dfrac{m'_2 - m}{m'_2} \times 100$ tends to be considerably larger than in the data of Greenwood and Woods. It also tends to be larger than in the other studies we have examined. This difference between Newbold's data and those of the other investigators is due in part to the fact that the mean numbers of accidents per person are rather large as compared to those of most of the other distributions. The irregu-

larly variable factors should become relatively less and less important in the long run. Still, this is not the whole explanation. The ratios computed for Newbold's material remain large even when compared to ratios from distributions with similar means. Table 6 presents these ratios as computed from the statistics given in Newbold's paper; the number of cases and mean numbers of accidents as given by Newbold and the variances (squares of Newbold's standard deviations) are also given. The figures may be compared to the corresponding ones in Table 5.

The median percentages are 71.6 and 56.05 for the men and women respectively. The range is very great, the largest figure being 90% while at the other extreme there is an obtained variance which is actually slightly smaller than that of the corresponding Poisson distribution; this distribution closely approximates a simple chance distribution. These percentages do not accurately represent the share of differences in accident liability in the variance of accident records in all cases. Inspection of Newbold's curves suggests that many of the obtained accident distributions deviate significantly from composite Poisson distributions. This matter was only partially investigated. The amount of work involved in the computation of composite Poisson distributions for thirty nine sets of data would have been prohibitive, particularly because these data are given by Newbold in the form of graphs rather than tables. Many of these graphs appear to have been inaccurately drawn, inasmuch as there are discrepancies between the totals of workers and accidents as read off from the graphs and as given in Newbold's Table.

Nevertheless, it can be shown that in some of Newbold's sets of data composite Poisson distributions are appropriate and the percentage of the variance attributable to differences in accident liability is large. As an example, Table 7[5] presents the data from Newbold's graph AIII, together with the corresponding composite Poisson figures. The closeness of the fit is apparent. The accident liability share is 75.8%.

Some of Newbold's sets of data suggest that the distribution of accident liability was a discontinuous one; in these sets of data the great bulk of the cases fit either a simple or a composite Poisson distribution, but there are also a few deviant cases which lie outside of such distributions. Most of the obtained variance of accident records due to accident liability may be due to the presence of these deviant cases; in other words, large deviations from the average accident liability appear only in a very small minority of cases. Thus Newbold's set EIII is essentially a distribution of the simple Poisson type, plus one markedly deviant worker. Set EV may be viewed as a distribution of the composite Poisson type (excess variance = 41%) plus 9 deviant workers. Table 8 presents these data.

[5] See footnote 2.

Table 6

Analysis of Newbold's Data

Newbold's Table No.	Number of Cases	Mean (m)	Variance (m_2)	$\dfrac{m_2' - m}{m_2'} \times 100$
EIII	226	.18	.59	68.0
FI	22	.27	.20	—
EIV	256	.41	.60	40.5
FII	81	.43	.45	4.2
MIV	106	.48	.59	19.1
EII	281	.51	.94	43.8
BII	299	.57	.81	29.6
I	190	.68	1.72	60.5
P	50	1.04	1.99	48.7
GI	47	1.47	3.76	60.9
GII	82	1.61	5.66	71.6
BI	148	1.81	5.11	64.6
MVI	218	1.95	7.13	72.0
MIII	181	2.50	6.60	62.1
AIII	304	2.56	10.56	75.8
EVI	93	2.66	12.53	78.8
EV	77	2.73	18.84	85.0
N	284	2.90	23.33	87.6
EI	440	3.64	13.76	73.1
AII	352	3.78	17.14	77.9
MI	301	3.94	14.90	73.3
MII	376	3.98	14.06	72.7
MV	92	4.07	18.15	78.6
EVII	57	5.60	56.25	90.0
AI	204	6.44	41.86	84.4
M1	380	.37	.53	30.6
GII	50	.52	1.04	50.0
GI	120	.63	1.64	61.5
MV	110	.65	1.46	53.5
I	161	.70	1.21	42.1
H	346	.79	1.35	41.3
MIII	142	1.06	1.77	40.1
BI	145	1.06	2.04	48.2
K	125	1.34	3.24	58.6
DII	98	1.39	3.39	58.9
BII	100	2.12	5.57	61.9
MII	161	2.30	8.58	73.2
C	58	2.43	7.88	68.7
DI	28	5.43	15.52	65.0

The differences between Newbold's findings and those of Greenwood and Woods, and of other investigators whose material is examined in this paper may possibly be attributed to the fact that her material consisted almost entirely of minor accidents. In spite of Newbold's statement,

the reporting of accidents may not have been complete. It is difficult
to ascertain the degree of completeness with which minor accidents were
reported and there may have been individual differences in the reporting
of accidents, producing the illusion of large differences in accident lia-
bility. On the other hand, constant personal characteristics may play
a more direct role in the causation of minor accidents than in that of

Table 8
Comparison of Two of Newbold's Sets of Data with Theoretically
Computed Distributions

Accidents per Man	Set EIII*		Set EV	
	Actual	Equal Liability Distribution (omitting 1 case)	Obtained	Unequal Liability Distribution (Composite Poisson) (omitting 9 cases)
0	201	197	24	22
1	21	26	22	19
2	2	2	8	12
3	1	0	5	7
4	0	0	6	4
5	0	0	1	2
6	0	0	1	1
7	0	0	1	0
8	0	0	3	0
9	0	0	2	0
10	1	0	1	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0
15	0	0	1	0
16	0	0	1	0
Total	226	225	76	67

* The discrepancy between the "actual" and the "equal liability" accident totals is
due to a similar discrepancy between the totals as given in a table in Newbold's paper,
and as obtained from her curve.

major accidents. Psychoanalysts generally believe that many accidents
are unconscious self-injuries. It is possible that such unconscious self-
injuries usually result in minor damage, just as in hysteria, in which minor
self-injuries are common while major injuries are unusual. Minor acci-
dents in industry may be often due to psychological mechanisms of the
hysterical type.[6]

[6] This hypothesis was suggested to the writers by E. Emmons.

The distribution of the Connecticut licensed car drivers is essentially a composite Poisson distribution. A Greenwood and Yule "unequal liability" distribution fits the data rather well, except at the upper end. The results can be accounted for if one assumes that the distribution of accident liability deviates slightly from a Pearson III curve. The estimated portion of the variance of accident records attributable to differences in accident liability is 21.2%. Table 9 presents these data.[7]

There is corroborative information from other sources, indicating that differences in accident liability often account for only a relatively small portion of the variance of accident records. The correlations between accident records in different periods of time reported by Newbold (9) and more recently by Ghiselli and Brown (5) are in most instances not high. Newbold's correlations range from −.01 to +.71, with a median of +.36. Ghiselli and Brown's correlations range from +.15 to +.80 with a median of +.42; we omit the intercorrelations between different kinds of accidents presented in both papers which are considerably lower. Such correlations justify inferences which are similar to those we arrived at by the use of a different method.

It should also be noted that the differences between automobile insurance rates for people with different accident records are nonexistent. This practice is in conformity with our findings. The usual textbook discussions of accident proneness would suggest very different insurance rates for different accident records.

When no composite Poisson distribution conforms to a set of accident data the suggested procedure is not applicable. The existence of factors must be assumed, which alter the shapes of the constituent Poisson curves. Changes in accident liability of people as a function of previous accidents encountered suggest a possible explanation of such results. We did not attempt to verify this possibility inasmuch as there seemed to be no way of arriving at a reasonably plausible hypothesis about the course of these changes in terms of information available at present. The only hypothesis suggested so far in the literature seems to have been the one implied in Greenwood and Yule's "Biassed distribution," and it is untenable theoretically and therefore unsuitable for research. This distribution is simply a Poisson distribution with a different first term. If there were no initial differences in accident liability, but the first accident changed the accident liability of its participants which would subsequently remain constant, the resulting distribution would not be an incomplete Poisson distribution, because the one accident class would not grow as in the "simple chance" case. An incomplete Poisson distribution could be produced only by continuing changes in accident liability with

[7] See footnote 2.

successive accidents, and it would be a strange coincidence if these changes should be so graded as to produce a tail end of a Poisson distribution which has a completely different derivation.

The distributions which deviate significantly from any composite Poisson distributions are two of Greenwood and Woods' distributions (their Table 1A and 1B), the distributions of taxicab accidents and the distribution of street car accidents. Inspection of the data indicates that the obtained distributions are more leptokurtic than Poisson distributions, and compounding several of the latter can often flatten out the resulting shape. Several of Newbold's distributions may be in the same category; they were not examined in detail for reasons stated earlier. The share of differences in accident liability in the total variance cannot be determined in such cases. The existence of other factors than differences in accident liability and unpredictable "chance" factors must be assumed.

Discussion

It must be remembered that not all differences in accident liability are differences in accident proneness viewed as an individual characteristic. This point is not a new one; it has been made among others by Newbold and by Cobb. It is disregarded by investigators who combine data about street car accidents or taxi accidents from different cities. In factory work, different jobs differ in conditions of safety. In automobile or other vehicle driving, the safety conditions are not necessarily the the same from route to route, in city compared with city. The amount of mileage driven, necessary driving in adverse weather, etc., must contribute more opportunities for accidents and these are not functions of accident proneness defined as an individual trait. For example, only 21.2% of the variance of the accident records of the Connecticut drivers was due to differences in constant accident rates. When one considers the hazards of driving just mentioned, it seems logical to state that there is not much room for differences in accident proneness as a psychological characteristic, insofar as these data are concerned.

We have pointed out that in many instances the portion of the variance of accident records attributable to differences in all forms of accident liability is relatively small as compared to the residual variance attributable to the operation of factors which are not predictable in terms of either the constant characteristics of people or of their previous accident records. These unpredictable or "chance" factors when operating alone give a so-called simple chance or equal liability or Poisson distribution. The expression "chance factors" should not be misunderstood. They are not necessarily unpredictable in terms of changing features of the life situation. Thus a well known psychoanalyst spoke to one of the writers about a man he knew who had a temporary period of accident proneness

as a result of marital trouble, during which time he had several near-accidents in rapid succession. "Chance" refers only to lack of predictability in terms of constant characteristics of the individual.

There are many kinds of such "chance" factors. One kind does not seem to have received enough attention in the literature. Even when a person is clearly at fault in causing an accident, the accident might not have occurred if the circumstances had been different. One of the writers was once in a car driven by a man who did badly enough to have caused a very serious accident: the driver became frightened by a wasp on his leg and stopped looking at the road; shortly afterwards the car travelled into a ditch at the bottom of an embankment to the left of the highway. There had been no cars in the other traffic lane at the moment he crossed it, the embankment was not steep and there was no accident. About half a mile further there was a steep drop into a river on the left side. The expression "luck" seems to be quite appropriate here.

As Cobb pointed out, the correlation between accident records and a perfect test of accident proneness need not be high. One cannot use any arbitrary criterion for classifying people as excessively accident-prone. For example, Poffenberger (10) states that "accident prone drivers are those who have two or three times as many accidents as the average driver . . . the term need not be restricted to auto accidents . . . for it covers equally well accident repeaters in industry." In many distributions examined here the number of accidents per person is one-half an accident or less. According to Poffenberger then, this would mean that persons with one or more accidents are to be considered as accident prone. This is obviously unfair. It is legitimate to select for study those people who have more than the average number of accidents but they should not be automatically classified as excessively accident prone without further evidence. Actually within a simple chance distribution some people are likely to have two to three times as many accidents as the average person. One can verify this by referring to the Poisson distributions in our tables. In most published distributions only a very small minority have accident records which lie completely above the point at which the Poisson distribution gives negligible values. As one approaches this point, one finds additional cases of more than average accident proneness, but some people with only average accident proneness who have had bad luck or temporary difficulties are also included in the group of people who have had many accidents. The problem of the exact estimation of the relative number of accident-prone individuals and bad luck individuals in any particular group of accident records is complicated. One should not attempt to make rough estimates without a comparison of obtained frequencies with the corresponding Poisson frequencies.

Summary

1. A commonly used method of comparing percentages of men and of accidents proves nothing about the existence of differences in accident proneness. Examples proving the inconclusive nature of the method are cited.

2. Comparison of obtained accident distributions with simple chance (Poisson) distributions establishes that there are differences in accident liability but does not indicate whether these differences are large or small and does not exclude the simultaneous operation of unpredictable "chance" factors.

3. Different accident records do not necessarily represent different degrees of accident liability. A method for analysis of the variances of accident records of people into two component variances is suggested, one component attributable to differences in accident liability, the other to unpredictable "chance factors." It is pointed out that the method is only applicable when the obtained distribution resembles a composite of Poisson distributions.

4. A number of published distributions of accidents are examined by the use of the above method. The variance attributable to differences in accident liability varies considerably.

In the distributions which are examined in this paper and which do not involve primarily minor accidents, the variance attributable to differences in accident liability is in most cases between twenty and forty per cent of the total variance of accident records. Although differences in accident liability should not be overlooked as a factor in the different accident records of people, the effect of this factor is rather small as compared to the residual 60 to 80 per cent attributable to unpredictable factors. It is therefore apparent that in many instances personal accident proneness, which is but one of the components of accident liability, has been an overemphasized factor.

References

1. Cobb, P. W.   The limit of usefulness of accident rate as a measure of accident proneness.   *J. appl. Psychol.*, 1940, **24**, 154–159.
2. DeSilva, H. R.   *Why we have automobile accidents.*   New York: John Wiley & Sons, 1942.
3. Elderton, W. P.   *Frequency curves and correlation.*   London: Layton, 1906.
4. Fisher, R. A.   *Statistical methods for research workers.*   London: Oliver & Boyd, 1948.
5. Ghiselli, E. E., and Brown, C. W.   Accident proneness among street car motormen and motor coach operators.   *J. appl. Psychol.*, 1948, **32**, 20–23.

6. Greenwood, M., and Woods, H. M. The incidence of industrial accidents upon individuals with specific reference to multiple accidents. *Industr. Fatigue Res. Bd., Rpt.* 4, 1919.

7. Greenwood, M., and Yule, C. V. An enquiry into the nature of frequency distributions representative of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Statist. Soc.,* 1920, 83, p. 255–279.

8. Kendall, M. G. *The advanced theory of statistics.* Philadelphia: J. B. Lippincott, 1943.

9. Newbold, E. M. A contribution to the study of the human factor in the causation of accidents. *Ind. Fatigue Res. Bd., Rpt.* 34, 1926.

10. Poffenberger, A. T. *Principles of applied psychology.* New York: D. Appleton-Century Co., 1942.

11. *Preventing taxicab accidents.* Metropolitan Life Insurance Company, New York, 1931.

12. Slocombe, C. S., and Brakeman, E. E. Psychological tests and accident proneness. *Brit. J. Psychol.,* 1930, 26, 29–38.

13. Viteles, M. S. *Industrial psychology.* New York: W. W. Norton & Co., 1932.

Method of Paired Comparisons and a Specification Scoring Key in the Evaluation of Jobs *

G. A. Satter

University of Michigan

Within recent years, public and industrial employers have increasingly attempted to place their wage structures on objective bases. Among the techniques employed to this end are those which are commonly referred to as "job evaluation methods." Collectively, these methods represent attempts to rate jobs in order to determine their relative worth with respect to other jobs and to use the job's standing, within the group of which it is a member, as a basis for assigning a dollars-and-cents value to it.

The most widely used methods fall into four general classes: (a) Those in which the operation of evaluation is one of comparing job against job in terms of the job's overall worth (Ranking Method); (b) in which it is one of comparing job against job in terms of specific "elements" or traits (Factor Comparison Method); (c) in which it is one of comparing the job against an arbitrarily defined scale of overall worth (Classification Method); and (d) in which it is one of comparing job against arbitrarily defined scales covering individual job traits or "elements" (Point Evaluation Method).

From time to time, various authors have described alternatives to, or modifications of, the above basic methods but for the most part these methods have retained their popularity with surprisingly few modifications. Thus, Viteles (10) and, more recently, Otis and Leukart (7) have recommended that the Method of Paired Comparisons be used as an alternate to the Ranking Method. So far as the present writer knows, no organization has ever given this recommendation a trial. Similarily, there are other scaling methods which might profitably be applied to the problem of jobs; on its face, the problem of scaling jobs does not seem to be pronouncedly different from that of scaling other subject matters. These alternative methods, too, have been neglected.

The present report describes the results of applying two psychometric techniques to the problem of building job scales in two industrial plants.

One procedure involves the application of the Method of Paired Comparisons and the other, the development of scoring keys which can be applied to job specifications. Both procedures are oriented toward developing a scale the points of which are defined by jobs and which can be used in making the kinds of job measurements which are helpful in setting up wage schedules.

I. Construction and Characteristics of Job Scales Built by the Method of Paired Comparisons

The Jobs Studied. The investigations reported here were carried out on the clerical jobs of two comparatively large, Midwestern, paper mills. In one mill (Plant A), 70 jobs supplied the subject matter of study; in the other (Plant B), 33 were studied. The group of 103 jobs covered a wide range of clerical skills; within the population were included Messenger and Mail Boy, the clerk classifications of accounting, purchasing, sales, billing, and scheduling departments, the specialized jobs associated with the operation of electric punched card equipment, and the supervisory jobs immediately associated with the jobs mentioned above.

The Job Analysis. In both of the plants, preparatory to the scaling project, all of the jobs were subjected to intensive study. The study methods were modeled after those developed and used by the United States Employment Service (11). In each plant, trained job analysts, working down the organizational chart, interviewed and observed to the organizational level of the jobs under study. The data collected thus represented the joint opinions of the employee performing the job, the supervisors immediately responsible for the job, the departmental head under whose jurisdiction the job fell, and the job analyst whose responsibilities were those of collecting, collating, and organizing the data into a formalized job description. The job descriptions were reproduced in final form only after the employees and supervisors who supplied the original data were given an opportunity to review and then to endorse them. In both plants, the completed job descriptions [1] were assembled in bound form, and in this form they served as the raw materials on which the judgments called for by the scaling operation were made.

The Collection of the Scale Data. In both plants the judgments called for by the Method of Paired Comparisons were made by those persons within the organizational structure who were presumed to know the jobs in question best, namely, the personnel working at them and the supervisors immediately responsible for them. In Plant A, thirteen judges (7 working on the jobs and 6 supervising them) and in Plant B (5 and 5) constituted a "scaling committee." The members of these two committees were called together for an orientation meeting by their respective industrial relations departments. At these meetings, the objectives of the project were outlined, the procedures to be used were reviewed, and the members of the committees were given the materials which they were to use in arriving at and reporting their judgments. These materials

---

[1] These job descriptions contained considerably more detail than one conventionally finds in the descriptions prepared for a job evaluation project. The objective in each case was to provide the reader, even though he had little previous contact with the job, with enough detail to permit a judgment of the skills and knowledges which it required, the responsibilities which it entailed, and the conditions under which it was typically performed—in short, to arrive at judgments concerning those characteristics which are conventionally associated with job worth.

consisted of a bound volume of the job descriptions, which had been prepared earlier, and a set of forms on which their judgments were to be recorded. It was then possible for the members of the committee to proceed independently, and at their leisure, to make their judgments. It might be pointed out here, that the routines of the Method of Paired Comparisons are particularly well adapted for use in the industrial situation. Since comparatively naive judges can be introduced to the task called for by the method with a minimum of training, it is possible to work with large numbers of judges who can proceed independently under a minimum of supervision.

By following this general procedure, the jobs in the two plants were scaled independently on four traits or "elements" which a preliminary review of the literature of job evaluation indicated as being potentially most useful in discriminating between clerical jobs. For purposes of the evaluation, these traits were defined in the following manner:

(a) *Educational Skills.* The degree to which the job demands preparatory skills (verbal, quantitative, etc.) which are most generally acquired in the schoolroom.

(b) *Work Skills.* The degree to which the job demands specialized skills which can only be acquired either through job training or by extended experience on the job.

(c) *Application Skills.* The degree to which the job makes special demands on the individual worker; the degree to which the job is unpleasant, tiresome, monotonous, dirty, etc.

(d) *Social and Personal Skills.* The degree to which the job requires human-relations skills—skill in supervising and in coordinating the activities of others.

Thus, both groups of judges were required to make their inter-job comparisons in four frames of reference. If the Method of Paired Comparisons had been used in its traditional form, this would have meant that each judge in Plant A would have had to make $\left(\dfrac{n(n-1)}{2} \right)$ 4 judgments (9,660) and those in Plant B, 2,112. To reduce the number of pairs of jobs in Plant A to a more feasible number, a suggestion which Uhrbrock and Richardson (9) made earlier was followed. By using key jobs, against which all comparisons were made, and groups of ten jobs in which only the in-group comparisons were made, the total number of judgments made by each judge was reduced from 9,660 to 3,660. These job groups were set up in the following manner. The investigating staff, selected from the group of 70 jobs, ten which in their opinion seemed to fulfill the dual criterion of being generally well known and which collectively represented the entire range of abilities required by the seventy. These constituted the "key group." [2] The sixty remaining jobs, then, were assigned to groups of ten in a random fashion. In preparing the worksheets for the judges, a scrambled order of pairs was used; each job title was presented first in half of the pairs; and the pairs involving the key jobs were interlaced throughout the whole list. No judge was informed that the key job device was being employed. In Plant B, the judges' worksheets called for the complete set of 2,112 judgments. Apparently, as we shall see later when the results from the two plants are compared, the modified procedure employed in Plant A did not distort the final results. Making and recording the judgments required from six to ten hours of the judge's time.

[2] The key jobs were: Dark Room Technician, Junior Stenographer, Mail Boy, Payroll Clerk, Record Clerk, Scheduling Clerk, Secretarial Assistant, Statistical Supervisor, Stenographer, and Telephone Operator.

Computation of the Scale Values. The data from each group of judges were summarized and, following a "shortcut" procedure recommended by Guilford (2), the scale value equivalents of each job were computed. This procedure was employed in preference to that called for by Thurstone's Case V of the Law of Comparative Judgment for the following reasons: (a) The small number of judges used hardly warranted the laborious operation of computing the several estimates of each scale separation which is required by the Thurstone procedure and (b) Guilford (2) has demonstrated empirically the comparability of scale values derived from using his abbreviated procedure and those derived from Case V procedure.

Results

The results of the operations described above were four skill scales which were presumed to be capable of measuring the dimensions on which wages for clerical jobs are commonly paid. At this point, the problem of combining the measurements yielded by these scales arises. If a suitable criterion is available, multiple correlational procedures are probably most appropriate. In a certain sense, the validation of job scales presents an even more difficult problem than is typically encountered in the validation of employee selection instruments; here, the problem is not only one of *measuring* the criterion, but, in the first place, of defining one. Lacking more suitable standards, in the typical wage evaluation project, job measurements are evaluated in terms of how well they reproduce the existing wage structure in the plant or the wage structures of other similar plants in the area. Both procedures obviously have serious shortcomings.

In the project described here, wage survey data for similar jobs outside the plant were assembled with the expectation that these data might be employed as a "criterion." Preliminary tabulations made it quite obvious that these data were incapable of generating correlation with anything, even themselves; the differences in wages paid for what were presumed to be similar jobs were often times as large, or even larger, than those which existed between different jobs. Accordingly, in both studies, the plants' prevailing rates were used as criteria in combining the scales values of the four skill scales.

A multiple regression equation was written for predicting rates from scale values. The multiple R's resulting from the application of the regression equation were .77 and .83 in Plants A and B respectively. In both plants, the Work Skills Scale contributed the most toward accounting for the total variance of "going rates." Apparently then, the kinds of measurements made by paired comparisons can yield measurements which are capable of ordering jobs with respect to their worth. The results reproduced in Table 1 also reveal that even better scales might be developed; the skill scales obviously do not measure independent

dimensions. This would suggest: (a) that the original choice of the traits was a poor one; (b) that the traits were poorly defined; and/or (c) that the judges were not highly proficient in making the kinds of discriminations which this project called for.

Table 1

Intercorrelations of the Scale Values Derived for Each of Four Job Traits and Their Correlations with Rates

Trait	Plant A*				Plant B*			
	2	3	4	5	2	3	4	5
1. Educational Skills	.93	−.49	.73	.71	.89	−.37	.74	.73
2. Work Skills		−.39	.75	.71		−.40	.82	.82
3. Application Skills			−.34	−.14			−.57	−.42
4. Social and Per. Skills				.66				.70
5. Going Rates								

* None of the inter-plant differences in the z-equivalents of the r's attain statistical significance.

Other characteristics of the scale values derived here may be pointed out. For one, the analyses suggest that these values are in general independent of the particular population of jobs chosen, i.e., that they have general validity. The correlations between the scale values of jobs in Plant A and those for jobs in Plant B, which the job analysis data indicated as similar in content, are presented in Table 2. These findings should be of special interest since they suggest that "standard scales" are feasible—that scales can be developed which will be of general applicability in job evaluation projects.

Table 2

Correlations between the Scale Values Derived in Plant A with Values Derived for Twenty-three Similar Jobs in Plant B

Job Trait	r_{AB}
Educational Skills	92
Work Skills	92
Application Skills	34
Social and Personal Skills	91

Further analyses of these data suggest high consistency in the judgments made by the several judges. Table 3 summarizes these findings. The coefficients reported in Column r_{11} are average intercorrelations between judges (5) and may be regarded as estimates of the reliability of the

Table 3

Reliability of the Judgments on which the Scale Values were Based

Jon Trait	Plant A		Plant B	
	r_{II}	r_{AA}	r_{II}	r_{AA}
Educational Skills	805	982	911	993
Work Skills	777	978	911	990
Application Skills	623	956	826	979
Social and Personal Skills	812	983	904	989

Table 4

Correlation between the *Sum of Judgments Made by Employee and Management Representatives*

Job Trait	Plant A r's	Plant B r's
Educational Skills	96	99
Work Skills	93	98
Application Skills	92	94
Social and Personal Skills	95	91

individual judge's judgments; those in the r_{AA} Column are the estimates resulting from applying the Spearman-Brown prophesy formula to the r_{II} values. Using comparatively large groups of evaluators obviously results in highly reliable judgments. These coefficients compare quite favorably with the few that are reported for "point-evaluation" judgments in the literature (4, 6). Further, from the above it may be presumed that the individuals who constituted the scaling committee were quite homogeneous in their outlooks toward the jobs which they evaluated —this, in spite of the fact that the committee membership was chosen to represent both employee and supervisory points of view. The correlations between the sums of employee and management judgments are reported in Table 4. This finding would suggest, then, that the attitudes of the judges who participate in a job scaling project (if we can assume that there were differences in the attitudes of the members of our groups) are not likely to color their judgments of the jobs. This finding is consistent with the findings of other investigators (1, 3) who have studied the scale values assigned to opinion statements by judges who differ pronouncedly in their attitude toward the object being investigated.

Summary: The Method of Paired Comparisons

In two investigations jobs were scaled on four traits by using the Method of Paired Comparisons. The results of these investigations

indicate that jobs can be scaled on these dimensions and that the measurements yielded by such scales can effectively be used to order jobs in a fashion which is valid for rate setting. The findings further suggest that the method used results in scale values which are independent of the particular population of jobs chosen.

At the practical level, the methods employed are particularly well adapted for industrial usage: (a) They permit the participation of large numbers of evaluators; (b) they can be employed with comparatively naive evaluators i.e., little training time is demanded; (c) even untrained evaluators report little difficulty in making the judgments called for; (d) the judgments can be made with a minimum of supervision and follow-up review; and (e) the resulting measurements are highly reliable.

II. Construction and Use of a Scoring Key for a Job Specification Form

In the same plants in which the investigations described above were made, trained job analysts collected and summarized other job data; they prepared specification forms which in form and content were somewhat like the *Worker Characteristics Form* employed earlier by the United States Employment Service in its job studies (11). The items, of which there were eighteen, covered various aspects of the skills and knowledges required by the jobs analyzed.[3] Each item was prefaced by a brief statement defining a particular skill (or knowledge) and this was followed by three or four alternative phrases or statements descriptive of various degrees of skill. These alternatives were drawn up arbitrarily to definite approximately equal distances along the skill scale. The following is a sample item:

Responsibilities for planning and laying-out work.

.... a. All work planned and laid out by the supervisor.
.... b. Particular class of tasks allocated to worker; lays out own schedule according to established routines.
.... c. Works on a job basis but has the responsibility for setting up own work operations and schedule.
.... d. Particular class of tasks allocated to worker; responsible for setting up own work operations and schedule.

Collection of the Data. As in the case of the job description preparation, described in Section I, the ratings called for by the specification forms were made on a cooperative basis by the job analyst, the immediate supervisor, and by the employee performing the job. One hundred and three such forms (70 in Plant A and 33 in Plant B) were prepared. These data supply the basis for the analysis reported in this section.

Analysis of the Data. Collectively, the items of the job specification form cover the same subject matter that was dealt with in the two scaling projects described above, so it seemed reasonable to presume that the ratings reported

---

[3] See Table 4.

on the specification sheets might be turned to the same usage as the paired comparisons data—namely, to order jobs with respect to their worth. Accordingly, a scoring key was developed for these items.

Each of the 18 ratings for the 70 jobs in Plant A was correlated with "going rates" and an equation was written for combining the "scores" of the individual items. In this equation, the individual item ratings were weighted in terms of their correlation with rates and the reciprocals of their respective standard deviations.

Table 5

Correlations between the Items of the Specification Form and Going Rates and the Standard Deviations of the Individual Item Ratings

Item on Specification Sheet	r	S.D.
1. Formal schooling demanded by the job	.46	.60
2. Skill in the use of numbers and numerical operations	.55	.75
3. Skill in the use of words—spelling and vocabulary	.29	.81
4. Skill in reading	.29	.99
5. Vocational training needed for the acquisition of job skill	.08	1.94
6. Training on the job	.61	1.72
7. Kind of supervision received on the job	.37	.91
8. Responsibility for planning and laying out work	.72	.82
9. Responsibility for making decisions	.58	.40
10. Conditions under which work is performed	−.03	.39
11. General nature of work—interesting, stimulating or routine and dull	−.28	.41
12. Physical demands of the job	−.04	.39
13. Supervision given to other workers	.52	.53
14. Relationships with other workers on the job	.32	.51
15. Relationships with persons outside the department	.46	.96
16. Skill in oral expression	.46	.72
17. Ability to maintain confidences	.31	.50
18. Appearance and dress requirements	.10	.35

Results

As a check on the accuracy of the scoring key developed (i.e., the ability of the key to reproduce the criterion on which it was built), the 70 specification forms were scored and the resulting scores correlated with rates. The coefficient was .89. Undoubtedly with further statistical manipulation of the item weights a larger proportion of the criterion variance could have been accounted for. The operation of correlating scores with the criterion on which the scoring key was originally built is, or course, no check of either the validity of the procedure nor its general usefulness. Accordingly, a similar set of specifications, which was developed in Plant B by another group of job analysts, was scored with the key developed in Plant A; again the resulting scores were correlated

with rates.   In Plant B, specifications correlated .92[4] with rates.   Thus, when an independent criterion and a new population of jobs is employed, the scoring key is found to be quite satisfactory.

Summary: A Scoring Key for a Job Specification Form

The procedures used in the development of a scoring key for job specifications forms has been described.   Such a scoring key was found to yield scores which are related to wage payments made to clerical workers.   There is some evidence to support the conclusion that such a scoring key developed in one plant may be of general usefulness in evaluating similar jobs in other plants.

Discussion

The two approaches to job measurement described here may be compared and contrasted.   As indicated above, they yield results which are very similar so that one's choice between them would probably be governed by considerations other than one of accuracy or validity of measurement.   First, it might be pointed out, the scoring key resulting from the application of Method Two, can be developed in a much shorter period of time primarily because the volume of data dealt with is much smaller; in contrast, the Method of Paired Comparisons, even when "short cuts" are employed, is always cumbersome.   Further, with Method Two, once adequate job analysis data have been collected, it is a comparatively simple task to collect the judgments called for by the job specification; but, it must be borne in mind that judgments of this sort can only be made by persons who have very intimate contacts with the jobs for which they are preparing specifications.   Training of the evaluators might, of course, overcome this limitation.

It might be argued then, that the Method of Paired Comparisons is more suitable for those kinds of projects where: (a) it is desirable to make the scaling project a cooperative one with comparatively large judging groups representing all interests, and (b) where one has a minimum amount of time to devote to the training of the judging group.   Apart from the fact that paired comparisons data are generally highly reliable, and that the method has a well-established theoretical basis, the above characteristics, in many industrial plants, would strongly recommend this method.

---

[4] Note that this coefficient is slightly higher than the one obtained on the initial check validation.   The difference in these two values does not attain statistical significance.

On the other hand, it is the writer's opinion, that the scoring-key method may be particularily valuable in certain special circumstances. Once such devices have been developed, they may be of particular usefulness in those situations where a comparatively small number of new jobs needs to be slotted into an already established wage structure. Or, again, where the manufacturing unit is so small as to make other more elaborate procedures impractical. The scoring-key method can easily be used as a supplement to any of the commonly used job evaluation schemes.

References

1. Ferguson, L. W. The influence of individual attitudes on construction of an attitude scale. *J. soc. Psychol.*, 1935, **6**, 115–117.
2. Guilford, J. P. The method of paired comparisons as a psychometric method. *Psychol. Rev.*, 1928, **35**, 494–506.
3. Hinckley, E. D. The influence of individual opinion on the construction of an attitude scale. *J. soc. Psychol.*, 1932, **3**, 283–296.
4. Jones, Alice M. Job evaluation of non-academic work at the University of Illinois. *J. appl. Psychol.*, 1948, **32**, 15–19.
5. Kelley, T. L. *Statistical method.* New York: Macmillan, 1923.
6. Lawshe, C. H., and Wilson, R. R. Studies in job evaluation. 6. The reliability of two point rating systems. *J. appl. Psychol.*, 1947, 31, 355–365.
7. Otis, J. L., and Leukart, R. H. *Job evaluation.* New York: Prentice-Hall, 1948.
8. Thurstone, L. L. A law of comparative judgment. *Psychol. Rev.*, 1927, **34**, 273–286.
9. Uhrbrock, R. S., and Richardson, M. W. Item analysis. *Person. J.*, 1933, **12**, 141–154.
10. Viteles, M. S. A psychologist looks at job evaluation. *Personnel*, 1941, 17, 165–176.
11. *Training and reference manual for job analysis.* Prepared by the Division of Occupational Analysis, War Manpower Commission. Washington: U. S. Gov. Print. Office, 1944.

The Effect of Equating Interest Test Items for Prestige Value

Elizabeth Fehrer

Brooklyn College

and

Hans Strupp

U.N.R.R.A.

A number of the currently used vocational interest scales require the subject to choose between pairs or groups of occupational titles or activities.[1] The assumption is made that degree of interest in a given field may be measured by the frequency with which one selects items that fall in this field over the other items with which they are grouped.

Interest scores obtained in this way should be most reliable when the items are matched for all factors that might influence choice except interests or personal values. Where choices are required between occupational titles, for example, preferences might at times be determined by such factors as the prestige of the occupations or their monetary return rather than by interest in a general type of work. Thus, if an item requires a choice between the occupations of United States Senator and scientific laboratory assistant, a person of high scientific interests might choose Senator because of its far greater prestige value. At the present time, however, there is no experimental evidence of the effect that factors such as these exert on interest scores.[2] It is the purpose of the present study to determine the extent to which the factor of prestige can influence such scores. The study originated in an attempted revision of the Allport-Vernon *Study of Values* (1). One type of item in the proposed revision consisted of pairs of occupational titles, the occupations being chosen to represent Spranger's value categories. From these items a person's score for a value category was to be determined by the frequency with which the occupations representing the category were preferred over those with which they were paired. It was in connection with the construction of this part of the test that the question arose concerning the

[1] See, for example, the *Kuder Preference Record* (4), the *Thurstone Interest Schedule* (11) and the *Occupational Interest Inventory* (Lee and Thorpe (5).

[2] The advisability of holding such factors constant, however, has been recognized. For example, in the construction of the *Occupational Interest Inventory* (5), the activities in each item have been roughly matched for job level.

necessity of matching the occupations for prestige value. The experiment to be described is an attempt to answer this question.[3]

The plan of the study involved the following steps:

1. The selection of occupational titles that fall into the Spranger categories.
2. The scaling of these occupations for prestige value by Thurstone's method of equal-appearing intervals.
3. The construction of an interest inventory in which the items consisted of pairs of occupational titles. From this scale, three separate scores could be computed for each Spranger value category. For example, one of the aesthetic interest scores was to be based on items in which aesthetic occupations were paired with other occupations equal to the aesthetic in prestige value. The second aesthetic score was to be computed from items in which the aesthetic occupations were higher in prestige value than the occupations with which they were paired. The third aesthetic score was to be computed from items in which the aesthetic occupations were lower in prestige value than the occupations with which they were paired. If prestige influences occupational preferences, then these three scores should differ significantly.

The only factors systematically varied in this study were the prestige values of the occupational titles and the interest categories in which they belonged. Other factors that might affect choice, such as financial returns or social service value, were neither isolated nor controlled. It is assumed that the influence of such factors would be similar to the factor of prestige, as Anderson (2) has shown high correlations between these factors.

Method of Selecting the Occupational Titles

The primary concern in the selection of the occupational titles was that they could be unambiguously classified into the Spranger value categories. Five interest categories were used: theoretical, economic, aesthetic, political, and social-religious. We decided arbitrarily to consolidate the social and religious values because (1) it was impossible to find a sufficient number of distinct occupations that fitted in the religious category and (2) both social and religious occupations seemed to involve humanistic and social-service interests and activities.[4]

---

[3] It is well known that in certain situations prestige is an important determiner of choice. The pioneer study of Moore (8) demonstrated that students' preferences for grammatical expressions, ethical situations and musical dissonances were influenced by knowledge of expert opinion. Studies by Marple (7), Sherif (9) and others have confirmed Moore's findings. In these studies, the opinion of experts was made explicit in the experimental procedure by assigning one of the choices or statements to the authority in question. In the present study, the factor of prestige operates in an entirely different way as it is inherent in the item.

[4] Correlational and factorial studies of scales measuring the Spranger values have yielded marked differences in the correlations between social and religious value scores. VanDusen, Wemberly and Mosier (12) report a correlation of .61; Ferguson, Humphreys and Strong (3) a correlation of .22. Lurie (6) reports a factor with high loadings on both scales. Even though social and religious values may be somewhat distinct, occupations that fall in the religious category seem to entail social service activities as well as high interest in spiritual values.

In searching for titles that could be classified in this manner, it soon became apparent that the list would have to be limited to professional, sub-professional and business occupations. These are the only types of occupations that seem to represent the Spranger values in an unambiguous manner. Skilled, semi-skilled, clerical and many other types of occupation had to be excluded. For example, the profession of artist seems clearly to fall into the aesthetic category whereas the occupation of house painter clearly fits neither the artistic, economic, nor any of the other categories. Consequently, the prestige range covered by the occupations chosen represents only a small fraction of the entire occupational prestige range. The range, however, is representative of that covered in certain existing interest scales.

One hundred occupational titles were selected, 20 to represent each of the five value categories.

Method of Construction of the Psychophysical Occupational Prestige Scale

Thurstone's method of equal-appearing intervals was followed in scaling the 100 occupational titles in respect to prestige value.

Fifty students in an advanced undergraduate class in experimental psychology served as judges. The 100 titles were printed on separate cards, arranged in random order, numbered, and a set was presented to each judge with the instructions to sort them into seven piles with apparently equal intervals between them. A seven-step scale was used instead of the traditional eleven-step scale since it was believed that with occupations as homogeneous in respect to social prestige as the ones chosen it would be impossible to discriminate eleven steps. With this exception, Thurstone's procedure was followed in determining the median and Q values for each occupational title. These values are presented in the second and third columns of Table 1.

Table 1

High, Median, Low and Mean Scale Values of the Occupational Titles in each Interest Category *

Interest Category	Scale Values			
	High	Median	Low	Mean
Political	0.60	1.91	6.35	2.45
Economic	1.75	4.45	6.45	4.43
Theoretical	1.72	3.34	5.22	3.37
Aesthetic	1.78	3.52	6.24	3.77
Social-Religious	2.25	3.86	5.70	4.01

* To reduce printing costs, Table 1 is presented here in greatly abbreviated form. The complete table, showing median scale values, Qs, and per cent unambiguous agreement in classification into value categories for each of the 100 occupational titles, has been deposited with the American Documentation Institute. For the six pages involved, order Document 2624 from the American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting $0.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or $0.50 for photocopies (6 x 8 inches) readable without optical aid.

Inspection of Table 1 shows rather marked differences in the prestige values of the occupations representing the various Spranger values. On the whole, the political occupations ranked very high in prestige. The nine highest ranking occupations belonged in this category. The means of the median scale values for the occupations in each category are as follows: political, 2.45; theoretical, 3.37; aesthetic, 3.77; social-religious, 4.01; economic, 4.43. These values of course refer only to the occupations selected in this study.

Since the political occupations ranked so very high in respect to prestige, this category had, later, to be excluded. There were simply not enough low-ranking political occupations to pair with others in constructing the interest scale.

The Q values (double the usual semi-interquartile range) of the occupations ranged form 0.5 for prime minister to 2.7 for poet. The mean ambiguity value was 1.72. These Q values are large compared with those found in the construction of an attitude scale. The size of Q is undoubtedly a function of the homogeneity of the items, but at the same time it also represents the fact that in our culture, professional and business occupations do not fall into a strict hierarchy in respect to prestige.

In the construction of the interest scale, occupations were not discarded on the basis of high Q values. The justification for this procedure was that the scores for each interest were to be based on a fairly large number of items. Consequently, individual differences in susceptibility to the prestige of the individual items should cancel.

Method of Checking the Accuracy of Classification of the Occupational Titles

In order to check on the accuracy of the classification of the 100 occupational titles into the Spranger value categories, 50 advanced undergraduate students were asked to sort them into the five categories. They were given a list of the 100 titles arranged in random order, descriptions of the five interest or value types adapted from Vernon and Allport (13) and they were asked to classify each title in one of the interest categories whenever this was possible. Such classifications will be referred to as unambiguous classifications. If an item seemed to fit into several of the categories, it could be placed in each. In such instances, the student was asked to indicate whether it seemed to fit equally well into both categories, or whether its placement into one seemed somewhat more suitable than the other. Such classifications will be referred to as coordinate classifications. These results are presented in Table 1 in the fourth column which shows the per cent of subjects placing the occupations in the designated value category.

The results indicated that there was fairly close agreenemt among the subjects concerning the proper classification of the majority of the occupational titles.

Eighty of the occupations were placed in the same interest category by at least 80% of the judges. Of these 80 occupations, 67 were unambiguously placed in the same category by at least 80% of the raters. That is, at least 80% of these raters indicated no coordinate category for these 67 occupations. Thirteen additional occupations were placed in the same interest category by at least 80% of the judges, but a small proportion of the raters indicated a second coordinate category into which the occupation might also fit. These 13 occupations are designated by an asterisk. The remaining 20 occupational titles yielded less than 80% agreement and were therefore eliminated in the construction of the interest scale. These 20 occupations are designated by a double asterisk.

The eighty occupations that met the criterion of 80% agreement in classification were distributed as follows among the five value categories: 16 theoretical, 19 economic, 17 aesthetic, 14 political and 14 social-religious.

Method of Constructing the Interest Scale

The two preliminary steps provided facts concerning the prestige values of the occupational titles and the degree to which each fitted the modified Spranger categories. The next and major task was to construct an interest inventory from which it would be possible to determine whether the prestige value of an occupation is a factor that will influence interest scores.

After eliminating the occupational titles in the political category and those titles that did not meet the criterion of 80% agreement in classification by the judges, 66 titles were available for constructing the scale.

The completed inventory was composed of 120 items, each item consisting of two occupational titles. Sixty of the items contained titles of equal prestige value, equal prestige value being defined as a difference of less than 0.50 points on the seven-point prestige scale. The mean difference in prestige value in these items was 0.19 points. The remaining 60 items contained titles that differed from each other by .60 points or more in prestige value. The mean discrepancy for these items was 1.53 points.

The inventory was constructed in such a way that the following four scores could be computed for each interest category:

1. *An Equal Score.* This score was derived from the 60 items in which the prestige values of the occupations making up the items differed by less than 0.50 points on the 7-point prestige scale. In this part of the scale, each interest category was compared with each other category tou times. That is, for example, 10 items involved comparisons of T and E titles; 10 involved comparisons of T and A titles; 10 involved comparisons of T and SR titles, etc. The maximum possible equal score for an interest category was 30.

2. *A Favored Score.* This score was derived from 30 of the 60 items in which the prestige values of the occupations differed by more than .60 points on the prestige scale. Here each interest category was compared with every other 5 times. The maximum possible favored score was therefore 15.

3. *The Opposed Scores.* These scores were derived from the remaining 30 items in the same manner as the favored scores except that here prestige operated against the selection of the occupations in an interest category.

4. *The Unequal Scores.* The unequal score was the sum of the favored and opposed scores, and represents the score for an interest category derived from the 60 items in which the prestige values of the titles differed. The maximum possible unequal score for an interest is 30. If prestige influences occupational preferences, the unequal interest scores should be more alike than the comparable interest scores derived from items in which the prestige factor is held constant.

Administration of the Interest Scale

The 120 items were arranged in random order, mimeographed, and the inventory was administered to 275 students in first-year psychology classes. Of the completed inventories, 180 were chosen for analysis, 90 from men and 90 from women students. For each student the 16 scores that have been described were determined, namely: 1. T, E, A and SR equal scores; 2. T, E, A and SR favored scores; 3. T, E, A and SR opposed scores; and 4. T, E, A and SR unequal scores.

Analysis of Interest Scale Results

Three types of analysis were undertaken to determine the effect of prestige value on occupational choices. The results of the three analyses are entirely consistent and all three show that prestige has no effect whatever on the interest scores.

1. The first type of analysis involved computing correlations between the equal, favored, opposed, and unequal scores for each interest category in order to determine whether scores based on the various types of items are in agreement. These correlations are shown in Table 2. They have been computed separately for men and women.

Table 2
Correlations Between the Equal, Favored, Opposed and Unequal Scores for Each Interest Category

Scores	Theoretical		Economic		Aesthetic		Social-Rel	
	Men	Women	Men	Women	Men	Women	Men	Women
Equal and Unequal	.93	.89	.95	.90	.87	.87	.87	.84
Equal and Favored	.82	.86	.92	.86	.82	.82	.68	.68
Equal and Opposed	.85	.84	.89	.84	.66	.79	.81	.78
Favored and Opposed	.62	.77	.80	.76	.50	.67	.47	.47

Scores on the equal and unequal scales are highly consistent. For men, the equal and unequal T scores correlate .93, the E scores, .95; the A scores .87 and SR scores .87. The corresponding correlations for women are .89, .90, .87 and .84. Scores on these scales are based on 30 items.

It is apparent that these interest scores are highly consistent whether based on items in which prestige cannot contaminate the scores or on items in which prestige might exert some influence on vocational choice. These high correlations also indicate high reliability for the scales.

The correlations between the favored and opposed scores are lower, ranging from .47 to .80. It must be remembered that these scores are based on only 15 items.

The method of correlation will show only whether scores on two scales are similar in rank order. It does not indicate whether one set of scores is numerically higher than the other. In order to determine whether the various scores for an interest category were directly comparable, a second type of analysis was undertaken.

2. The second method of analysis consisted of directly comparing the interest scores based on the equal, favored, opposed and unequal scales. To facilitate comparison, the raw scores were first converted into percents.[5] These percent scores then represent the proportion of times that titles in the interest category under investigation were preferred to the titles with which they were compared. These scores are presented in Table 3. The scores for men and women are presented separately as the two sexes differed in their preferences.

Table 3
Equal, Favored, Opposed and Unequal Percent Scores for Each Interest Category

Scores	Theoretical		Economic		Aesthetic		Social-Rel	
	Men	Women	Men	Women	Men	Women	Men	Women
Equal	43	36	60	31	49	66	48	67
Favored	44	36	65	33	46	58	48	55
Opposed	39	39	60	37	49	70	49	73
Unequal	42	37	63	35	47	64	49	64

It is evident that the factor of prestige has no significant effect on these percent scores. For the 90 men, for example, we find that the T titles are chosen 43% of the time from the equal scale, 44% of the time from the favored scale, and 39% of the time from the opposed scale. None of the differences are significant.

Again, the SR titles are chosen 48% of the time from the equal scale, 48% of the time from the favored scale and 49% of the time from the opposed scale. Again, the differences, this time in the opposite direction, are not significant. The results for women are comparable.

[5] It should be remembered that the maximum raw equal and unequal scores were 30 whereas the maximum raw favored and opposed scores were 15.

The large majority of the differences are not statistically significant. Of the 32 critical ratios presented in Table 4, only three are significant and here the differences are not in the expected direction. For example, there is a significant difference (C. R. equals 4.18) between the equal and favored SR scores for women. Reference to Table 3, however, shows that the equal percent score is 67 and therefore higher than the favored score of 55.

Table 4

Critical Ratios Between Equal, Favored, Opposed and Unequal Scores for Each Interest Category

Scores	Theoretical		Economic		Aesthetic		Social-Rel	
	Men	Women	Men	Women	Men	Women	Men	Women
Equal and Unequal	0.56	0.48	0.69	1.23	0.50	0.76	0.12	1.20
Equal and Favored	0.25	0.03	1.29	0.66	0.83	2.43	0.18	4.18
Equal and Opposed	1.19	0.87	0.07	1.75	0.02	1.07	0.37	2.10
Favored and Opposed	1.39	0.83	1.22	1.05	0.84	3.48	0.53	6.32

The differences in the four obtained scores for each interest category cannot be attributed to the factor of prestige as there is no general tendency for the favored scores to be higher than the equal scores nor are these generally higher than the opposed scores. Differences in the obtained scores are presumably due to (1) chance factors and (2) the particular titles that occur in the equal, favored, and opposed scale items. Although no analysis has been made of this last factor, it is obvious that certain occupations are generally more popular than others. This factor was not controlled in the construction of the scale.

3. The third type of analysis consisted in computing for each individual a susceptibility to prestige score. This score was obtained from the 60 items in which the prestige values of the occupations composing an item differed. The score consists simply of the number of favored occupations selected minus the number of opposed occupations. The maximum possible range of this susceptibility to prestige score is from minus 60 to plus 60, a positive score representing susceptibility to prestige. The median susceptibility to prestige score for men was zero; for women, minus 6. It is obvious that there is no tendency to favor occupations high in prestige.

Summary

The results of these three analyses seem to show clearly that, insofar as college students are concerned, preferences for occupations within the range studied here are not determined by the prestige which is accorded

to that occupation. Although differences in prestige are recognized, as is demonstrated by the fact that occupations can be scaled for this variable, occupational preferences are not determined by this factor. Instead, preferences are apparently determined by far more basic interest patterns, and, as the consistencies of the scales suggest, these interest patterns constitute a relatively stable component of the personality. This, of course, has often been pointed out by Strong (10).

It may, in addition, be safe to conclude that in the construction of professional interest scales when occupational titles make up the items, the prestige values of the items may be ignored.

Received October 8, 1948.

References

1. Allport, G. W., and Vernon, P. E. *A study of values.* Boston: Houghton Mifflin & Co., 1930.
2. Anderson, W. A. Occupational attitudes of college men. *J. soc. Psychol.*, 1934, 5, 435–465.
3. Ferguson, L. W., Humphreys, L. G., and Strong, F. W. A factorial analysis of interests and values. *J. educ. Psychol.*, 1941, 32, 197–204.
4. Kuder, G. F. *Preference record.* Chicago: Science Research Associates, 1942.
5. Lee, E. A., and Thorpe, L. P. *Occupational interest inventory, advanced series. Manual of directions.* Los Angeles: California Test Bureau, 1943.
6. Lurie, W. A. A study of Spranger's value-types by the method of factor analysis. *J. soc. Psychol.*, 1937, 8, 17–37.
7. Marple, C. H. The comparative susceptibility of three age levels to the suggestion of group versus expert opinion. *J. soc. Psychol.*, 1933, 4, 176–186.
8. Moore, H. T. The comparative influence of majority and expert opinion. *Amer. J. Psychol.*, 1921, 32, 16–20.
9. Sherif, M. An experimental study of stereotypes. *J. abnorm. soc. Psychol.*, 1935, 29, 371–375.
10. Strong, E. K. *Vocational interests of men and women.* Stanford Univ. Press, 1943.
11. Thurstone, L. L. *Thurstone interest schedule.* New York: The Psychological Corporation, 1947.
12. Van Dusen, A. C., Wimberly, S., and Mosier, C. Standardization of a values inventory. *J. educ. Psychol.*, 1939, 30, 53–62.
13. Vernon, P. E., and Allport, G. W. A test for personal values. *J. abnorm. soc. Psychol.*, 1931, 26, 231–248.

Personal Preference Differences among Occupational Groups

Mary F. Mosier

Bureau of Naval Personnel, Navy Department, Washington, D. C.

and

G. Frederic Kuder

Duke University

This study was conducted in order to explore the occupational patterns resulting from the use of a new measure of preference developed by one of the authors. The mean scores of twenty occupational groups, ranging from unskilled labor groups to highly professionalized vocations, were compared with the average scores of a group of unselected men. There has also been investigation of the differences in mean scores among three occupational levels. The three occupational levels are approximately those included in the major occupational groups of the *Dictionary of Occupational Titles* (1), as 0—Professional and Managerial occupations, 1—Clerical and Sales occupations, 4 through 7, Skilled and Semi-skilled occupations. Results reported here are to be considered as suggestive, rather than conclusive, since an earlier abbreviated, and less reliable, form of the test was used in making the study.

The new measure of preferences, entitled *Preference Record—Personal*, consists of five scales. Each of the scales has been developed so as to have high correlations among the items comprising a scale, and low correlations among the scales. The content of the scales may be described as follows:

- A. Preference for taking the lead and being in the center of activities involving people.
- B. Preference for dealing with practical problems and everyday affairs rather than interest in imaginary or glamorous activities.
- C. Preference for thinking, philosophizing, and speculating.
- D. Preference for pleasant and smooth personal relations which are free from conflict.
- E. Preference for activities involving the use of authority and power.

It may be noted that the scales are based on recorded preferences. There is no implication intended that the scales measure actual facility in the areas described.

The sample of unselected men is composed of the first 450 respondents to the random sampling as described in the Manual (2). The only criterion for inclusion in this group was that the test blank be filled out in accordance with the instructions. For the sake of convenience we shall hereafter refer to this population of unselected adult males as the "base group."

Our occupational samples have been drawn from an additional group of more than 1000 returned *Preference* blanks obtained through the original sampling referred to above. Information as to the vocation of the subject was obtained from the personal data section of the test blank wherein he was requested not only to name but also describe his work. This double requirement of the subject that he both name and describe his employment made it possible to check doubtful titles and descriptions against those of the *Dictionary of Occupational Titles*, and thus increased the validity of our occupational groupings. For example, when job descriptions were checked we frequently found the title "Accountant" self-conferred on a "Bookkeeper."

The count and classification of the occupations represented in our sample of more than 1000 revealed more than fifty different occupations reported by our subjects. It was decided that those occupations represented by twenty or more cases could be analyzed for the purposes of exploration. Accordingly, the test blanks of the members of the twenty such occupations were isolated for study. Table 1 lists the occupations grouped according to the appropriate level. There were 577 cases represented in the twenty occupations which could be grouped as follows: Professional and Managerial, 10 occupational groups with a total of 298 cases; Clerical and Sales, 5 occupational groups and 130 cases; and Skilled and Semi-Skilled, 5 occupations and 149 cases. Table 1 gives the number of cases for each occupational group.

Procedure

Mean scores and standard deviations for the base group of unselected adult males were computed for each of the five scales of the *Preference Record—Personal.* For each occupational group, only means were computed since, for purposes of testing the significance of mean differences, the variance of the unselected group was considered a much better estimate of the population variance than the variances obtained from the comparatively small occupational groups. Comparisons were then made between the means for a particular occupation and those of the base group. By such comparison we could observe the difference between the average member of an occupation and the average member of a general group with reference to the scale in question. In order to observe the effects

Table 1

Mean Scale Scores for Base Group, Occupations and Occupational Levels for
Preference Record—Personal

Group	N	A		B		C		D		E	
		$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Base Group	450	10.06	4.1	13.35	3.2	11.67	3.5	18.68	5.2	14.18	3.9
Professional and											
Managerial	298	10.1		13.4		12.1		19.8		15.0	
Accountants	35	10.0		13.0		12.1		19.1		15.1	
Bus. Managers	25	9.4		13.2		13.0(+)		17.8		16.4(+)	
Chemical Engrs.	21	9.9		14.7(+)		11.6		20.0		15.3	
Mechanical Engrs.	29	9.7		13.9		11.5		21.1(+)		14.7	
Office Managers	21	8.3(−)		13.2		13.0(+)		21.2(+)		15.0	
Personnel and Coun-											
seling Workers	23	11.6(+)		14.1		14.3(+)		20.1		16.4(+)	
Plant Managers	24	9.5		13.0		12.1		18.9		14.3	
Retail Managers	56	10.3		13.4		11.5		19.3		13.7	
Sales Managers	26	11.5(+)		11.8(−)		12.6		18.8		16.5(+)	
Teachers	38	10.0		13.9		11.1		21.4(+)		14.8	
Clerical and Sales	130	10.3		12.9		12.0		19.3		14.4	
Acct. Clerks and Tellers	35	8.7(−)		13.4		12.6		19.0		14.1	
Gen. Off. Clerks	22	9.6		13.1		11.1		20.4		13.6	
Insurance Slsmn.	23	11.8(+)		13.3		11.8		21.7(+)		14.4	
Salesmen other than to											
Consumer	27	12.8(+)		11.4(−)		11.7		18.9		14.9	
Salesmen to Consumer	23	8.7(−)		13.3		12.5		17.0		15.4	
Skilled and Semi-skilled											
Trades	149	9.9		13.6		11.3		18.4		13.4	
Carpenters	24	7.9(−)		14.3		10.2(−)		17.4		13.2	
Electricians	33	10.3		13.6		11.2		18.5		14.0	
Factory Workers	50	10.4		13.4		11.4		18.8		12.3(−)	
Foremen, Mfg.	20	11.6(+)		13.2		11.4		18.6		14.9	
Telephone Linemen	22	8.5(−)		13.4		12.4		18.1		13.7	

+ means significantly greater than the base group at the 5% level.
− means significantly lower than the base group at the 5% level.

of occupational level on test scores, we have also computed the mean scores for each of the occupational levels on the five scales, and studied the differences found between the three major groups. The purpose of this procedure was to determine the relation between enjoyment of a certain type of activity and the position in the occupational hierarchy. For example, are the average scores of professional people higher on the scale referring to preference for use of authority than those found for workers in the trades? Or, do clerical and sales employees indicate less

favorable attitudes toward taking the lead than do professional and managerial?

The significance of differences obtained was estimated in terms of the standard error of the difference between means. However, the standard deviation used in computing the standard error was that of the base group, rather than that of the occupational group in question, since this appeared to be a more accurate determination of the standard deviation of the universe than would any single sub-sample. The N in each such comparison was that of the occupational subsample.[1]

Results

The means and standard deviations for the base group sample are shown in Table 1. Also in Table 1 are shown the mean scores for the three occupational levels, (1) Professional & Managerial, (2) Clerical & Sales, (3) Skilled and Semi-Skilled Trades. These three levels were compared with each other rather than with the base group of unselected men. Inspection of Table 1 indicates substantial differences between the base group and one or another of the three occupational levels did occur on several scales. However, the significance of these differences has not been computed. In Table 2 we have listed differences between the mean scale score among the three occupational levels, and for those differences found to be above the 5% level, the critical ratio is shown in parentheses.

Discussion of Results

Any interpretation of results found in the study must be prefaced by a reminder of the small size of the occupational samples. Analysis of the data scale by scale yields information about both the scale and the attitude of the various occupations toward the activity embraced by the scale.

Scale A. The items on this scale relate to a preference for taking the lead and being in the center of activities involving people. Inspection of Table 1 shows that more significant differences between occupational groups and the general population sample occurred on this scale than any other. The mean score for ten occupational groups differed significantly from that of our base group sample. Five occupations indicate highly favorable attitudes toward the activities involving social leadership: (1) Personnel & Counseling Workers; (2) Sales Managers; (3) Insurance Salesmen; (4) Salesmen other than to Consumer; and (5) Foremen. The

---

[1] The formula used was: $SE_{Diff.} = \sqrt{\dfrac{\sigma_u{}^2}{450} + \dfrac{\sigma_u{}^2}{N}}$, where σ_u is the standard deviation of the base group.

five occupations showing less-than-average enjoyment in these activities were: (1) Office Managers; (2) Accounting Clerks & Tellers; (3) Salesmen to the Consumer; (4) Carpenters; and (5) Telephone Linemen. Two of these occupations which showed less than average interest in leading people or being in the center of a social situation are at first glance surprising. We refer to the Salesmen to Consumer group and the Office Manager group, both of which might be expected to indicate average or above-average enjoyment of people. An examination of the personal data and job descriptions of members of these two groups throws some light on their attitudes not indicated by their job titles. We find among the Salesmen to Consumers, a large number of individuals who report physical handicaps or old age or failure in some other line of work. A warranted generalization would seem to be that this group of salesmen did not select their vocation but were forced to it through inability to function in another profitable field. Therefore, it appears the Salesmen to Consumer group, comprising door-to-door salesmen and canvassers, and not retail store clerks (as the title might imply), are atypical of salesmen in general. Perusal of the data yielded by the group of Office Managers suggests that many of them are actually little engaged in working with people. Rather they list as their duties "ordering supplies," "checking incoming orders," "making work schedules," "reviewing reports," "coordinating." After checking these job descriptions there is less mystery in the responses these two groups have made on Scale A.

The three occupational levels do not differ significantly in their preference for taking the lead and being in the center of things involving people. Preference for these activities characterize skilled and semiskilled trades as much as professional and managerial occupations. That the skilled group scored as high as it did may be attributed to the significantly higher mean score for the 20 foremen.

Scale B. The content of this scale relates to a preference for activities of a practical nature, rather than imaginary or glamorous pursuits. When comparison is made between the various occupational groups and the base group, we find in Table 1 that there were three occupational groups showing significant differences. Chemical Engineers show a strong preference for practical activities, while Sales Managers and Salesmen other than to Consumer are less interested in these matters than is the average man. A number of occupations which might be expected to show high preference scores failed to do so for these samples. A glance at Table 2 shows that the group of occupations labelled "Skilled & Semi-Skilled Trades" have a mean score that is significantly higher than that of Clerical & Sales occupations. No other significant difference between the occupational level groups was found for Scale B.

Scale C. This scale may be described as preference for "thinking"—thinking of a philosophical or speculative nature. Significantly high mean scores were found for three occupations: Business Managers, Account Clerks and Tellers, and Personnel and Counseling Workers. The only significantly low mean was that for Carpenters. While the results for Personnel Workers and Carpenters may be in accordance with the hypothesis concerning the trait measured, those for the other two groups are not. Moreover, the lack of high scores for Chemical and Mechanical Engineers and Teachers, all of whom deal with abstract and conceptional ideas, does not seem consistent with the identification of this scale as thinking philosophizing, and speculating. More cases and data on additional occupations are needed before this trait can be definitely identified.

Table 2

Differences between Means of Three Occupational Levels on the Five Scales of
Preference Record—Personal *

Scale	Prof. and Man. Minus Cler. and Sales	Prof. and Man. Minus Sk. and Semi-Sk. Trades	Cler. and Sales Minus Sk. and Semi-Sk. Trades
A	−.193	.209	.402
B	.516	−.150	−.666 (1.95)
C	.103	.820 (2.37)	.717 (1.94)
D	.422	1.382 (2.67)	.960 (1.74)
E	.579	1.668 (4.32)	1.089 (2.65)

* Figures in parentheses represent the critical ratio of the significant differences found between means.

Table 2 indicates that preference for activities measured by Scale C is related to occupational level. We observe that Professionals are higher than Clericals, but this difference is not significant. However, both Professionals and Clericals are, on the average, more favorable toward these activities than is the average Trade worker, and this difference is found to be beyond chance expectations.

Scale D is designed to measure the individual's preference for activities of an agreeable nature—activities free from conflict. The occupational group showing the highest enjoyment in pursuits of an agreeable nature was Insurance Salesmen, consistent with the stereotype of this group as highly amicable. The next three groups, in order of mean score, were Teachers, Office Managers and Mechanical Engineers. No occupational group yielded a mean score significantly lower than that of the base broup.

We find here on Scale D differences among the occupational levels in

their mean scores. The average Professional is higher than the Clerical according to our figures, but the difference is small enough so that it can be attributed to chance factors. However, we find the critical ratio of the difference between Professionals and Trades of such magnitude as to be called very significant, and the differences between Clericals and Trades significant. We must assume that the average man working in the skilled or semi-skilled trades can be expected to be considerably less interested in activities of a pleasant, amicable nature than white collar or college-bred men.

Scale E's items relate to enjoyment of the use of authority and power. High mean scores that show important and real differences from the mean of the base group are to be found only in the Professional & Managerial Group. Business Managers, Personnel & Counseling Workers and Sales Managers say that they certainly do like to exercise power. Lawyers are being found to be significantly high on this scale in a study made after the one reported here. The only group which indicated less than average satisfaction in these activities were Factory Workers. It is perhaps appropriate to describe here the composition of the group "Factory Workers." These were respondents to the random sampling who stated they worked in a plant or manufacturing concern and who described their jobs by mentioning a single simple function repeatedly performed, such as one phase of an assembly procedure. The group is heterogeneous in that a great many different industries are represented, but they are similar in that their work was described as taking place in a factory where they performed a single mechanical or motor act. The degree of skill ranged from unskilled to semi-skilled.

Analysis of mean scores by occupational level indicates that the "trait" measured by Scale E is related to position on the occupational ladder. Professionals & Managerial Workers are higher than Clerical & Sales Workers, in general, in their preference for pursuits using authority. The critical ratio, however, is only 1.43. A difference as large as this in this direction could occur about 8% of the time through the operation of chance. But Professionals are sufficiently higher in their mean score than Trades that we can say the difference is too great to have arisen by chance more than once in a thousand. The difference between Clericals and Trades is significant at the 1% level.

The discussion of results up to this point has been from a "vertical" standpoint, i.e., the results have been examined in terms of the scales. A "horizontal" view of the results presents somewhat different information about the occupations and appears worthwhile.

There were five of the twenty occupational groups which showed no significant deviation from the mean of our base group on any one of the

five scales. These occupations were Accountants, Plant Managers, Retail Managers, General Office Clerks and Electricians. With reference to the qualities measured by the *Preference Record—Personal* these people, or occupational groups, appear to be typical of the general population. It is interesting to note that all three occupational levels are represented in this group of occupations which show the same characteristics as the general population with regard to the variables measured. A study of larger groups may, of course, reveal significant differences.

On the other hand, we find three occupations differed in their means from that of the base group on three of the five scales. These occupations showing atypical pattern were Personnel & Counseling Workers, Office Managers, and Sales Managers. These three occupations can be said to differ from the average man more than any of the other seventeen occupations with regard to the characteristics studied.

Of interest also are the results for the three occupational levels. We found no important differences between the two higher levels on any of the five scales. However, the Skilled and Semi-skilled group, as shown in Table 2, shows rather marked differences from the other two groups, Professional & Managerial and Clerical & Sales.

In comparing the highest occupational level with the lowest we see that the preferences of Professionals and Managers are higher than those of Skilled and Semi-skilled trades workers for activities involving "philosophical thinking" (Scale C), pleasant relations (Scale D) and the use of authority (Scale E).

Comparison of the group of Clerical & Sales occupations with those of the Skilled & Semi-skilled trades workers, reveals the same differences as those described in the paragraph above plus a difference in the opposite direction on Scale B. On Scale B, preference for activities of a practical nature, we observe in Table 2 that the trades occupations show a significantly higher mean score than the clerical group.

Summary

This study indicates that each of the five scales makes some discriminations by occupation, and that there is a relation between some occupations and the characteristic measured by each scale.

Fifteen occupations differed on one or more of the scales. Five of the occupations did not differ from a general population sample on any of the scales.

We found that differences also occurred that are related to the level of the occupation in the economic or educational scheme. These differences were rather large and significant between the group of occupations

labelled Skilled & Semi-Skilled Trades and the group, Professional & Managerial, and also the group, Clerical & Sales.

References

1. United States Employment Service, U. S. Department of Labor, *Dictionary of Occupational Titles, Part I.*
2. Kuder, G. Frederic, *Manual for Kuder Preference Record—Personal.* Chicago, Illinois. Science Research Associates. 1948.

The OL Key of the Strong Test and Drive at the Twelfth Grade Level *

Stanley R. Ostrom

Department of Public Instruction, Dover, Delaware

One of the baffling problems facing educators today is that of finding an instrument that will determine, with an acceptable degree of accuracy, which pupils possess the pattern of traits that enable them to make the best use of their abilities. If an instrument could be found that made it possible for a counselor to distinguish subjects whose backgrounds and native endowments were such that they could easily be activated to exert a maximum of energy from subjects whose backgrounds had pre-disposed a more lethargic set, it would be possible to predict scholastic and vocational success much more accurately than·is now the case.

The Occupational Level Key of the Strong Vocational Interest Blank for Men has been recommended by Strong (6, p. 195) and Darley (1, p. 60) as an instrument that will enable a counselor to make this distinction. Kendall (3) and Ostrom (4) have demonstrated that the OL key of the Strong Blank can be used with considerable confidence for this purpose at the College Freshman level. This paper reports an attempt to determine the utility of the OL key at the twelfth grade level.

Two hundred twelfth grade boys enrolled in four Central New York high schools formed the sample. One-half of these boys cooperated in an intensive study and the total group participated in a study which utilized their academic aptitude as measured by the American Council on Education Psychological Test scores, drive [1] as measured by the OL key, and four year academic grade averages.

The 100 boys who cooperated in the intensive study were selected in the following manner: from three of the four high schools a total of sixty boys were chosen so that twenty of them had very high scores on

[1] For purposes of simplicity, the pattern of traits discussed in the first paragraph will be represented in subsequent pages of this report by the term drive.

the OL key, twenty of them had very low scores, and twenty of them had scores that clustered around a scaled score of fifty. Thus, three groups which were differentiated by OL were obtained. In the fourth high school, forty boys were chosen in such a manner that their OL scores fell on a continuum. This was done to determine whether or not spuriously high relationships between OL and the experimental variables would be obtained in the other three high schools due to sampling methods.

Three new instruments were devised for purposes of checking on the results of the OL key. These instruments were: (1) a Teacher's Rating, (2) an "Open End" interview, and (3) a "Guess Who" questionnaire. The Teacher's Rating (see Figure 1) was produced to measure drive in the four following areas: (1) drive for hobby satisfaction, (2) drive for scholastic achievement, (3) drive for co-curricular achievement, and (4) drive for vocational attainment. Each of the four traits was measured on a seven-point scale with discrete descriptions utilized for each of the seven points.[2]

The second instrument, a "Guess Who" questionnaire (see Figure 2) was devised in an attempt to determine how the young men felt about the drive and persistence of their peers. In this instrument ten descriptions were listed with space provided where each subject could name the three of his peers who best signified the quality required of each statement.[3]

The third instrument was the interview (see Figure 3). The writer interviewed each individual, making use of the basic set of ten questions. The subject was permitted to elaborate on each question as much as he desired. From time to time, secondary questions were asked to encourage the subject to enlarge on the response given to the primary question. By means of the ten questions, the writer attempted to elicit from the subject information from his background, his past school, work, and hobby experiences as well as his hopes and plans that gave evidence of the presence or absence of drive.[4]

It was necessary to quantify the results of the three new instruments before they could be of any value in determining relationships between their results and those of OL.

The Teacher's Rating Scales were filled in by five teachers who had known each boy for at least one year. Each trait was measured on a seven-point scale, hence the maximum score obtainable on each trait

---

[2] $r = .89 \pm .03$, N = 40. Test-retest method with two week interval.

[3] $r = .94 \pm .02$, N = 40. Test-retest method with two week interval.

[4] No measure of reliability determined.

Fig. 1. Rating Scale

On this sheet you will find a scale on which you can rate four kinds of drive. Will you read each scale through carefully and copy the number that corresponds to your rating of each boy in the column that corresponds to the trait on the sheet that accompanies the scale. If you do not have sufficient information to rate all traits, leave those for which you have inadequate knowledge blank.

I. Drive for Hobby Satisfaction: Desire for achievement and success at hobbies

1. Has no hobby.
2. Joins hobby group more for social benefits than because of hobby. May at times indulge in a hobby if relatively popular but soon loses interest.
3. Has some interest in hobbies. Gets discouraged easily and drops a hobby if relatively minor obstacles arise.
4. Participates actively in hobbies. Enjoys working at them and will achieve success.
5. Is an active hobbyist on 1 or 2 hobbies but is interested in several. Will endure handicaps in order to work at his hobby.
6. Spends much of the time on hobbies. Persists at them after his friends have lost interest. Will overcome serious obstacles to achieve at his hobby.
7. Spends all his leisure time and spending money on hobbies. He has a hobby or hobbies that have persisted for a number of years. Hobbies will play a major role in his vocational future.

II. Drive for Scholastic Achievement: Desire to achieve success in school subjects

1. Appears to have no interest in scholastic situation. Even with much encouragement makes little progress in attempting to achieve in school subjects.
2. Will put forth some effort to do simple school tasks. Much urging by teachers is necessary if he is to exert himself.
3. Needs encouragement if he is to overcome serious obstacles. Hard tasks are usually solved through own motivation.
4. Will persist with arduous tasks if given some inspiration by teachers. Will achieve at a moderate level on his own.
5. Will solve difficult problems only with a great deal of prodding. Will surmount minor obstacles through own volition.
6. Will overcome arduous tasks through own motives. Seldom needs urging by teachers in working to capacity. Usually works beyond requirements of instructors.
7. Sets extremely difficult goals for self and perseveres until they are attained. Requires no supervision from teachers.

III. Drive for Co-curricular Achievement: Desire to achieve in school activities outside the realm of strictly curricular offerings

1. Does not participate in co-curricular activities.
2. Participates in a limited number of activities and will assume no responsibility. Continued participation is dependent upon constant outside pressure.
3. Participation is limited to group activities. Some encouragement is usually necessary to insure continued participation. Immediate success in activity may be sufficient to hold subject's interest.
4. Takes part in 1 or 2 school activities. Will endure moderate privations and minor initial failure experiences in order to insure success. Assumes responsibility for minor duties.
5. Exhibits enthusiasm for co-curricular activities. Willingly accepts responsibility. Limits voluntary participation to 1 or 2 activities. Will persevere in activities in which he is interested in face of fairly difficult obstacles.
6. Shows a dogged determination in the activities for which he has special interest. Will participate willingly and actively in others if asked. Exhibits a very high degree of responsibility.
7. Exerts a tremendous amount of energy in co-curricular school activities. Works with determination and enthusiasm in a number of school programs. Persists over serious obstacles.

IV. Drive for Vocational Attainment: Desire to attain a well paying and socially acceptable position. May involve power, prestige, service to others, contributions to society, or personal enhancement.

1. Appears to have no idea of what his vocational future may be. He is not motivated to press any interest in the matter. Seems satisfied to let nature take its course.
2. Has not given much thought to his vocational future. Appears to be willing to accept a very mundane form of vocational activity.
3. Plans and aspirations appear in haphazard activities. Will accept employment which does not challenge his capabilities.
4. Has expressed interest in work at the skilled trade or white collar level. Has a strong desire for security.
5. Has his sights set on technical or semi-professional standing in the world of work. May be interested in a small private enterprise.
6. Has well defined plans for training schedule leading to technical or professional standing. Desires to attain a position of leadership in his community.
7. Has set a very high occupational goal and is driving himself to achieve this end. Goal may be beyond his reach or very near the ceiling of his capabilities. Has a burning desire to "set the world on fire."

was seven, and the maximum score obtainable from all four traits was twenty-eight.

Teachers vary in the ratings they give in two ways. First, some tend to rate all students relatively high while others are more conservative in their evaluations. Second, some raters are very discriminating in rating students, and the results they obtain vary over a large portion of the range; others are much less discriminating, and the ratings they give cover only a small portion of the range.

FIG. 2.—"Guess Who."

Following you will find a number of descriptions which have been listed. You will also be given a list of boys from your class. We are asking you to list the three boys from this list that best fit each of the statements. You are not asked to choose only your friends. The boys who best fit the descriptions may be boys you do not like very well. Thank you for your cooperation.

1. The boy whom you feel will make the most of his abilities:
   1................................. 2................................. 3.................................

2. The boy you would like to have with you if you were lost in a blinding snowstorm:
   1................................. 2................................. 3.................................

3. The boy who will work the hardest to gain an education:
   1................................. 2................................. 3.................................

4. The boy who works hardest in school now:
   1................................. 2................................. 3.................................

5. The boy who participates most in co-curricular activities:
   1................................. 2................................. 3.................................

6. The boy who works hardest outside of school:
   1................................. 2................................. 3.................................

7. The boy whom you expect will become the most famous:
   1................................. 2................................. 3.................................

8. The boy on whom you would be most willing to bet in a boxing match if he were matched with a boy of equal size, strength, speed, and ability:
   1................................. 2................................. 3.................................

9. The boy who has to be knocked down the most often in a fight before he will quit:
   1................................. 2................................. 3.................................

10. The boy who would be most apt to come back and win in a set of tennis if the score against him were 5–4 with the count in the final game being "Add" against him:
    1................................. 2................................. 3.................................

Your name: ....

Fig. 2.   Questions Used for Personal Interview.

1. Would you mind telling me what your father does for a living?

2. Do you know the highest grade (or degree) your father and your mother attained?

3. Would you give us a fairly complete picture of your work experience?

4. What do you expect to do when you are through with school?

5. Would you discuss your plans for gaining the training required for that job?

6. Would you care to tell me how you got interested in ....?

7. Would you say that hobbies have played any part in determining your vocational goals? If so, how?

8. Do you feel that you are satisfied with your school progress?

9. Would you say that your school work to date is a fair indication of your abilities?

10. What would you like to be doing in ten years?

To correct for these two difficulties the ratings of all the participating teachers were converted into comparable measures.[5]

After the ratings had all been made comparable, the average rating was then changed to a T-score[6] (2, p. 99) so it could be utilized in further statistical procedures.

In using the "Guess Who" device the boys were asked to list three boys from the group in their school whom they felt best satisfied each of the ten descriptions. It was possible for a boy to list one of his peers on several questions. This happened on numerous occasions. The scores, which were obtained by counting the number of times each boy was listed, ranged from four to ninety-three. The scores thus obtained were also converted to T-scores.

The results of the interviews were quantified by the following method: the boy's responses from each question were rated form one to four in terms of their expression of drive. A response that denoted much drive

$$[5] \ X_{BA} = \left(\frac{\sigma_A}{\sigma_B} \right) X_B - \left[\left(\frac{\sigma_A}{\sigma_B} \right) M_B - M_A \right].$$

Where X_{BA} equals measurement in distribution B transformed into the terms of distribution A.

X_B equals original measurement in distribution B.

σ_A equals standard deviation of distribution A.

σ_B equals standard deviation of distribution B.

M_B equals mean of distribution B.

M_A equals mean of distribution A (2, p. 121).

Since seven ratings were used, the mean score for distribution A was taken as the middle score or four. The standard deviation was arbitrarily set at 1.8.

[6] For purposes of this study T-score is used in the sense of Walker's Z score, thus not assuming normality.

was rated one; a response that denoted very little drive was given a value of four. The *total* of the ratings for all questions comprised the score for the interview. The boy with the lowest score, fourteen, thus measured highest on drive in this measure. Table 1 shows the distribution of the four ratings for the 100 boys. These scores were also converted to T-scores but in a reverse manner so that low scores resulted in high T-scores.

It was possible to correlate the results of the three original instruments with OL scores since as has already been stated, the scores obtained through the three instruments were all changed to T-scores. The correlations are indicated in Table 2. It is evident from the table that OL correlates

Table 1

Distribution of Ratings Given the 100 Twelfth Grade Boys on the "Open End" Interview

Rating	Number of Times Rating Was Used	Per Cent of Total
1. (Highest Value)	134	14
2.	304	33
3.	322	35
4. (Lowest Value)	169	18
Total	929	100

Table 2

Relationship Between Three Variables and OL in a High School Population *

Variables	School I (N = 29)	School II (N = 16)	School III (N = 15)	School IV (N = 40)	Total (N = 100)
OL—Interview	.56±.13	.39±.23	.46±.22	.56±.11	.48±.08
OL—Teacher's Ratings	.54±.14	.60±.17	.24±.26	.28±.15	.41±.08
OL—Guess Who	.41±.16	.43±.22	.37±.24	.38±.14	.41±.08
Teacher's Rating Interview	.72±.09	.71±.13	.71±.14	.57±.11	.59±.06
Teacher's Ratings Guess Who	.74±.09	.73±.13	.55±.20	.56±.11	.61±.06
Interview—Guess Who	.57±.13	.38±.23	.43±.23	.30±.15	.39±.08
OL—Total T-Scores of the Instruments	.56±.13	.53±.19	.35±.25	.51±.12	.53±.07

* Spearman's Rank Difference formula was used in the first three schools due to the small number of students. In School IV and the Total, Pearson's Product-Moment formula was used.

to a very significant degree with each of the three instruments and that it correlates to a highly significant degree with the total score which resulted when the T-scores of each of the three instruments were added. It will also be noted that the magnitude of the correlations obtained in School IV does not vary significantly from those obtained in Schools I, II, and III. This tends to show that choosing three groups of high, average, and low OL scores, as was the case in School I, II, and III, did not in this instance permit spuriously high results.

As a further check, Chi Square was used to determine the relationship between OL scores and the scores obtained in the Teacher's Ratings, "Guess Who", interviews, and total ratings. As can be seen in Table 3 all four Chi Square results are of a magnitude that justify the rejection of the Null Hypothesis at the one percent level.

Table 3
Relationship Between OL and Three Variables *
for High School Population

Variables, Total 88	Chi Square	Confidence Level
Guess Who and OL	15.22	>1
Teacher Ratings and OL	24.06	>1
Interviews and OL	16.23	>1
Total and OL	22.12	>1

* The Null Hypothesis states that the three OL groups: high, average, and low, do not constitute different populations in terms of the "Guess Who" ratings, Teacher's Ratings, interview results, and the total results obtained by adding the T-scores of the three variables for each boy. A chi square of 13.277 was necessary to reject the Null Hypothesis at the 1% level of confidence.

Having found a relatively high relationship between OL and the instruments described above, an attempt was made to determine the relationship between OL and school achievement as measured by school academic grade averages.

The assumption on which the study was based was that excellence in school was to some extent determined by motivation or effort expended. To find this relationship the 200 boys from the four high schools were divided into two groups, the first being made up of boys with high OL and the second made up of boys with low OL. With these two groups the following two questions were posed: (1) do the two groups differ significantly in scholastic achievement? (2) if so, how much of this difference is due to OL?

Table 4 registered an F-ratio of 5.66 which was of a magnitude that places the confidence level for the rejection of the Null Hypothesis between the five and one per cent levels, thus answering the first question in a doubtful affirmative. To answer the second question it was necessary to adjust for the other variable, academic aptitude as measured by the

Table 4
Analysis of Variance of Honor-Point Ratios
N = 100 Twelfth Grade Boys

Source of Variance	Degrees of Freedom	Sum of Squares	Mean Square	F*	Test of Hypothesis**
Within	198	5417.12	27.35		
Between	1	154.88	154.88	5.66	Remain in doubt
Total	199	5572.00			

* Where F = greater mean square/lesser mean square. By referring to Snedecor's tables of F (5, 222–225), we may use the following three rules in testing the hypothesis: (a) reject the hypothesis tested, if the calculated value of F is greater than the 1% point given in the tables; (b) accept the hypothesis tested, if the calculated value of F is less than the 5% point given in the tables; (c) remain in doubt, if the calculated value of F lies between the 5% and 1% points given in the tables.

** The Hypothesis tested is a null hypothesis concerning the difference between means of groups, i.e., there is no significant difference between the means of groups. (The 1% point was 6.76 and the 5% point was 3.89.)

Table 5
Complete Analysis of Variance and Covariance—100 Twelfth Grade Boys *
(Partialling out the Effect of Academic Ability)

Source of Variance	Degrees of Freedom	Sum of Squares y^2	Sum of Square x^2	Sum of XY	DF	Adjusted or Reduced Sum of Squares	Mean Square	F	Test of Hypothesis
Within means of groups	198	5417.12	18722.70	4931.12	197	4118.35	22.02		
Between means of groups	1	154.88	1039.68	401.28	1	14.85	14.85	.7	Accept
Total	199	5572.0	19762.38	5332.40	198	4133.20			

* See footnotes for Table 4.

American Council on Education Psychological Examination. When the data were adjusted for academic aptitude by means of covariance, as shown in Table 5, an F-ratio of only .7 emerged. Since this was not of a magnitude to justify rejecting the Null Hypothesis, the answer to Question 2 must be that the difference in academic grade averages due to OL was almost negligible.

Summary

1. A definite relationship was demonstrated between OL on one hand, and out-of-school and co-curricular evidences of drive on the other hand. Thus it appears that boys who evidence much energy and activity in the less formal school situations and in everyday life situations as a rule give responses on the Strong Blank which result in high OL.

2. No relationship was demonstrated between OL and high school academic grade averages. The reasons for this can be only conjecture but a few of them are ventured. It might be that high school does not present a challenge to most boys with the result that marks which enable a boy to "get by" are satisfactory. The possibility that boys satisfy their desires to achieve through co-curricular activities and life situations cannot be ignored. Furthermore, it is common knowledge that high school marks are not always valid. Questionable marks could easily cause a relationship to fail to emerge. It might be pointed out further that the use of the Strong Blank among high school students is questionable due to the immaturity of high school students. Strong has pointed out that interest patterns change quite extensively during the high school years. He states "roughly speaking, one-third of the change in interests is between 15.5 and 16.5 years, one-third between 16.5 and 18.6 years, and one-third between 18.5 and 25 years (6, p. 259)."

References

1. Darley, J. G. *Clinical aspects and interpretation of the Strong Vocational Interest Blank.* New York: The Psychological Corporation, 1941.
2. Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill Book Company, 1942.
3. Kendall, W. E. The occupational level scale of the Strong Vocational Interest Blank for Men. *J. appl. Psychol.*, 1947, 31, 283–287.
4. Ostrom, S. R. The OL key of the Strong Vocational Interest Blank for Men and scholastic success at college freshman level. *J. appl. Psychol.*, 1949, 33, 51–54.
5. Snedecor, G. W. *Statistical methods.* Ames, Iowa: Collegiate Press, Inc., 1946.
6. Strong, E. K. *Vocational interests of men and women.* Stanford, California: Stanford University Press, 1943.

An Objective Evaluation of Counseling

Barbara A. Kirchheimer, David W. Axelrod, and

George X. Hickerson, Jr.

University of California Counseling Center, Berkeley

The development of objective criteria for evaluating the effectiveness of counseling has traditionally been a matter of extreme difficulty. Apparently, the first studies in the evaluation of faculty counseling are the 1925 unpublished studies of Paterson and Langlie at Minnesota; and that of Lemon (6) at Iowa in the same year dealing with counseling by professionally trained counselors. Lemon's work consisted of intensive remedial training for half of the lowest decile of students on the Iowa Qualifying Examination. At the end of three years, Holladay (5) summarizing Lemon's study reported that the "counseled" group were making a better academic adjustment than the equally weighted group left to their own devices. However, Freeman and Jones (4) in a final report of the same group state that at the end of their college career there was no difference between the two groups, because academic failure appeared later for the experimental group.

Use of the "spoon-feeding" type of counseling is shown in the studies of Cowley (3) with Ohio State Freshmen football players, Newland and Ackley (7) with high school sophomores and Williamson (9). In Williamson's study made on Art College students, he found as Paterson and Langlie had previously found with Engineering students, that the grade point average of probationary students was not improved by faculty counseling. Williamson concluded that grade point average is not adequate as a criterion of the effectiveness of counseling, or that other counseling methods must be used than those involved in his study.

Two years later, Williamson (10) showed significant increases in honor point ratio for a student group counseled by trained counselors at the University of Minnesota Testing Bureau when compared with a matched non-counseled group. In a later study Williamson and Bordin (13) made use of subjective evaluations of adjustment and cooperation, in addition to grade point average. In a further paper on this same study (11) the authors show that both adjustment and grade point average are significantly better for a counseled group than for a matched non-counseled group. Since both these criteria are significant at the 1% level, one wonders why it was felt necessary to go beyond the grade point

average and use the subjective composite criteria. In criticizing the techniques for evaluating counseling, Williamson and Bordin (12) feel grade point average is a poor criterion because of the dissimilarity in pattern of subjects taken. However, the alternative of using standardized achievement tests has limitations in comparing achievement in a number of areas. Moreover, the fallibility of the measuring instrument itself must be emphasized.

Blackwell (2) in a client-centered counseling program at the University of Texas reports significant increases in grade point average for a counseled group of 40 compared with a matched non-counseled group. Ward and Tyler (8) at the University of Oregon show a slightly better record for a counseled group than for a matched non-counseled group in grade point average, as well as on their special scale attempting to measure college adjustment. Beaumont (1) in a somewhat confusing article purports to show that discrepancies in academic adjustment were due in a large measure to differences in academic counseling. However, he points out the fact that most "academic" counseling is more concerned with subjugating the individual to the academic machine than with the integration of the individual's personality.

In most academic settings, grades alone are an objective indication of progress or adjustment. In view of the fact that grades are the only specific criterion of which we are in possession, that they lend themselves to objective treatment, and that, with all their weaknesses, they are the accepted gauge of academic success or failure, the present authors have adopted this criterion as the most workable measure so far available whereby to evaluate the success of a counseling program.

In evaluating the effect of counseling, an amplified approach might include considering the results upon grades of change of major course of study. A change of major often accompanies vocational and/or educational counseling, and the effect of such a marked step is insufficiently investigated. From comparison of pre- and post-counseled grades, we may have some clues to the effectiveness of the change of major itself, and of the professional counseling which produced it.

Selection of Groups and Methodology

Accordingly, it was decided to study veteran students at the University of California, Berkeley Campus. High admission requirements and fairly rigid disqualification regulations result in a rather homogeneous, high caliber population. The average grades of undergraduate veteran students appear in Table 2.

If any evaluation of counseling is to be made, the kind of counseling under investigation should be described. It is individual, consisting of

as many interviews as are required to develop *mutually* a vocational and/or educational plan, with all needed testing individually planned, and use of the occupational library maintained by the Occupational Information Specialist of the Center. The psychological training and experience of the Counselors and Psychometrists, and the services of a Consulting Psychiatrist insure that each counselee's total personality and situation is considered, rather than simple vocational or educational symptoms. Techniques are eclectic, with the constant objective of formulating an optimal, realistic plan. A coordinate objective is the growth of the counselee so that he may carry out the plan. Counseling is concerned with the development of the individual rather than with the improvement of grades. An important point is that the educational plan is a joint agreement between counselor and counselee. Counseling cannot be superimposed but must be the result of mutual understanding and *real acceptance*.

Because of the dual approach in determining effect on grade average of counseling, and of change of major, with and without counseling, the following groups were used:

I	*Counseled Change.*	Changed major as a result of a mutual decision of Counselee and Counselor.
II	*Non-Counseled Change.*	Changed major without any contact with the Counseling Center.
III	*Counseled No-Change.*	Continued same major as a result of a mutual decision of Counselee and Counselor.
IV	*Non-Counseled No-Change.*	Continued same major without any contact with the Counseling Center.

Counseled groups were selected from the files of the Counseling Center of the University of California, Berkeley, operating under contract with the Veterans Administration for advisement of veterans. All veteran students included in Groups I and III received counseling under Public Law 346 (G. I. Bill), coming voluntarily to seek assistance for a variety of reasons.

The University of California, Berkeley, Counseling Center has unfortunately been in existence only since October, 1946, and few cases of veterans who had changed major were available who had had at least one semester of enrollment prior to, and had completed one semester following, counseling. All records filed in chronological order were reviewed by a clerk with instructions to select every case of a student enrolled as an undergraduate in the University at the time of requesting counseling, who signified intention at the last counseling interview of enrolling in a different Department, School, or College within the University of California the following semester. No case meeting these requirements was eliminated.

There was considerable range in type of change. The only common change was from some form of Engineering (Mechanical, Electrical, Civil, and Industrial) to Business Administration, which accounted for six of the thirty-five

cases.   Examples of other changes were Chemistry to Architecture, Physics to Social Welfare, Forestry to Agricultural Economics, Chemistry to Psychology.

These students must have completed their counseling one full semester preceding the selection in order that grades following the change might be obtained.   It had been hoped that grades two semesters before and two semesters after counseling might be obtained, but this was not possible at this time. Williamson's ((11) interpretation of his data is that the effect of counseling is apparent in the first quarter following counseling, and no further increase in grades occurs in succeeding quarters.   Unfortunately his hypothesis cannot be explored with our groups at this time.

The number of our *Counseled Change Group*, for these reasons, was only 35, and for purposes of comparison other groups were constituted of equivalent size.   Various studies have matched groups on the basis of intelligence, sex, age, etc.   Williamson (12) points out that in such matching it is impossible to include such significant variables as motivation, personality, or emotional stability.   Matching on intelligence test scores might have been desirable, but no such scores were available for those groups which had not gone through the Counseling Center.   An unpublished study made at this Center by William R. MacKay compares the grade point average of those veteran students availing themselves of the Center's facilities with the general grade point average of all veteran students at the University.   This study showed no significant difference in grade point average of the counselee, and also that on the ACE Psychological Test, the average counselee score was at the 82.7 percentile (σ 20.16). As already pointed out, the high admission requirements and fairly rigid disqualification regulations result in a rather homogeneous high caliber population within the University.   The present authors, therefore, feel that a random sample drawn from the University population may be assumed to be roughly equivalent in intelligence and that no matching between groups was important other than that all subjects should be undergraduate male veteran students at the University of California, Berkeley.

Only six changes of major were made by members of the *Counseled Change Group* between Fall 1946 and Spring 1947, and 29 such changes occurred between Spring 1947 and Fall 1947.   For this reason, for the other groups the two semesters Spring 1947 and Fall 1947 were used for comparison, and are designated 1st and 2nd semester.

The *Non-Counseled Change Group* was, like all other groups, collected by a clerk, who reviewed University alphabetical records of veterans for these two semesters, selecting the first 35 males who registered a change of major between these semesters, and who had at no time contacted the Counseling Center.

The *Counseled No-Change Group* consisted of the first located 35 males in the Center's files who were enrolled in the University as undergraduate students for the two semesters in question, and who did not in this period change their majors.   All cases satisfying these criteria were retained.

The *Non-Counseled No-Change Group* was selected in the same manner as Group II, except that the first 35 cases enrolled both semesters who did not change majors and who had not contacted the Counseling Center comprised this group.

The college year distribution of the *Counseled Change Group* (Group I) is as follows: Freshmen 6; Sophomores 17; Juniors 11; Seniors 1.   The year distributions of the other 3 groups very closely approximated this, with very few students who were not divided between the Sophomore and Junior years.

Grade point average at the University of California is computed on the basis of:

> Three points per unit of credit for A,
> Two points per unit of credit for B,
> One point per unit of credit for C,
> No points per unit of credit for D and F.

The sum of the grade points divided by the number of units for which registered yields the grade point average (G.P.A.).

In the handling of our data, the significance of the differences obtained was calculated according to the formula for the critical ratio of the difference over standard error of the difference. A value of the critical ratio of 1.96 is reliable at the 5% level, and a value of 2.58 is reliable at the 1% level.

Results

The grade point averages for the two semesters studied for all four groups are given in Table 1.

Table 1

Summary of Grade Point Average and Changes

Number—35 Each	1st Semes. G.P.A.	σ 1	2nd Semes. G.P.A.	σ 2	G.P.A. Change	σ ch.	Range of G.P. Changes	C.R. of Change
I. Counseled Change	1.13	.61	1.65	.60	.52	.66	—.70 to 1.99	3.50
II. Non-Counseled Change	1.17	.62	1.41	.55	.24	.62	—.66 to 1.50	1.70
III. Counseled No-Change	1.46	.63	1.54	.56	.08	.49	—.70 to 1.01	.56
IV. Non-Counseled No-Change	1.46	.62	1.39	.63	—.07	.13	1.00 to .76	.47

From Table 1 it may be noted that the *Counseled Change Group* improved from slightly better than a C average (1.13) to a B — average (1.65), or a gain of .52 grade points, a change which is significant at better than the 1% level.

Since it was felt that grade point average may be affected by elective courses not pertinent to the major, a calculation was also made of only courses in or required by the major. Two cases were necessarily eliminated in this calculation because, although they had officially made a University transfer, they had not yet undertaken any courses in the new field. Incidentally, these two cases were among the seven whose grade point average was lowered following the statement of change. For the remaining thirty-three cases the mean grade point average in the major courses only before the change was .946 (a level of deficiency) and afterwards was 1.68, or an increase of .734 grade point on the average.

Twelve students of the 35 received a grade point average of less than 1.00 (deficient level) for their semester's work prior to counseling, while only one student received a grade point average of less than 1.00 after counseling. A number of individual examples may be cited. One student

who had a .77 grade point average (down grade points) with C's and D's, under a new major the following semester rated three A's and 1 B or an A — average (2.75). One student receiving 2 D's and 2 F's improved to four C's.

Whereas the *Non-Counseled Change Group* had only a slightly higher grade point average initially than the *Counseled Change Group*, its increase (.24) with a change of major was less than half as great, a change significant only at the 9% level.

The change in grade point average of the *Counseled No-Change Group* (.08) and the *Non-Counseled No-Change Group* (— .07), with Critical Ratios of less than 1, were not significant changes.

It had previously been found by the Coordinator of Veteran Affairs of the University of California, Berkeley Campus, that grades were inversely related to size of study load, i.e., number of units carried. For undergraduate students the averages were as follows:

Table 2

Grade Point Average Compared with Average Number of Units Carried

Semester	Study Load	Grade Point Average
Fall, 1945	11.2	1.91
Spring, 1946	12.5	1.57
Fall, 1946	13.5	1.53
Spring, 1947	14.1	1.37
Fall, 1947	14.2	1.41

It was felt, therefore, that such an increase as shown by Group I might be partially a result of a decreased study load. As can be seen in Table 3, the study load of the *Counseled Change Group* went up, and therefore, such explanation for their higher average must be rejected. The *Non-Counseled Change Group* on the other hand did decrease their study load slightly. However, both of the No-Change Groups were carrying a slightly heavier program in the second semester than were the change groups, but again the counseled group was slightly more heavily loaded than the non-counseled, although they decreased rather than increased their program.

As can be seen, the most significant difference, much beyond the 1% level, is between the *Counseled No-Change* and *Non-Counseled No-Change* Groups. The difference between the *Counseled Change* and *Non-Counseled Change* Groups is significant at the 7% level.

For purposes of comparison, groups were combined to increase their size. In Table 1 it is evident that both groups which did not change major have a higher initial grade point average than the groups which changed. Combining them, thus giving two groups with an N of 70 each, it is found that the *No-Change Group* (III and IV) has an initial grade point average of 1.46 σ .11, while the *Change Group* (II and I) has an initial average of 1.15 σ .11. The difference of .31 has a critical ratio of 17.51, significant at the 1% level, indicating that the students who changed majors, whether counseled or not, had a significantly lower grade point average initially than those who did not change. This may indicate that this group whose grades were below their potentialities endeavored to improve them by a change of major or by seeking counseling. Of course, half of the *No-Change Group* also sought counseling. This may also indicate that better grades are achieved by those in appropriate fields of study.

Table 3

Change in Average Study Load (units)

	1st Semester	2nd Semester	Change
I. Counseled Change	13.62	14.31	+.69
II. Non-Counseled Change	14.23	14.00	—.23
III. Counseled No-Change	14.77	14.70	—.07
IV. Non-Counseled No-Change	14.23	14.50	+.27

Table 4

Critical Ratio of Differences of Changes Between Groups

	II. Non-Couns. Change	III. Couns. No-Change	IV. Non-Couns. No-Change
I. Counseled Change	1.84	3.17	5.22
II. Non-Counseled Change		1.20	2.89
III. Counseled No-Change			5.55

By combining the groups according to whether counseled or not, we find the *Counseled Group* (I and III) makes an increase in grade point average of .30, σ .62, while the *Non-Counseled Group* (II and IV) makes an increase of .09 in grade points, σ .47, a difference of .21 grade points in favor of the counseled groups. This difference has a Critical Ratio of 2.31, significant at the 2% level.

Discussion

As has been mentioned, the groups in this study were necessarily small, and therefore the conclusions that may be drawn are limited. It is hoped that when possible this study will be repeated with a larger sample. Since the methodology of this study has afforded suggestive results it is also to be hoped that this study will be repeated with other populations.

We feel that the use of the criterion of grades is warranted in view of their importance for the survival of the student, his future opportunities for professional training, or for employment. We fully recognize, however, how few aspects of "counseling effectiveness" such a criterion may evaluate. It is a task of the future to develop criteria for these less objective areas. When this problem has been mastered, it may be found that additional criteria will show more clearly the value of vocational and educational counseling.

It is particularly apparent in this study that most students with academic deficiencies eradicated these deficiencies in the semester following counseling, regardless of whether they changed their major. These data imply the social value of counseling in the salvaging of deficient students. However, it is no less clear that students making satisfactory grades can benefit from counseling.

From the standpoint of evaluating counseling, we cannot, of course, generalize beyond the particular type of counseling under study. Without careful, intensive vocational and educational counseling on an individual basis, with concern for the individual as a whole, results may differ.

The improvement of grades by counseled students might be attributed to whatever factors differentiated those students seeking counseling from those who do not, a possibility considered in similar studies. With the inclusion of a group who changed majors without counseling, we feel that we have effected some equalization of whatever factors may lead students to take action of one kind or another to improve their situation. As shown in this study the improvement made by those who were counseled and changed major is considerably greater than that made by those who changed major independently. We cannot, of course, demonstrate conclusively that the scholastic improvement of the counseled groups as compared with the non-counseled groups was due to the counseling, since counseling itself is a complex of many variables. Such a possibility, must, however, be considered. The other studies, with similar counseling, have in general shown similar results.

Summary

1. A group of male veteran undergraduate students who changed their majors as a result of counseling improved their grade point average

significantly, despite an increase in number of units carried. Improvement is even more marked if only major subject course grades are considered.

2. The difference in grade point average improvement between two groups of male veteran undergraduate students who did not change their majors, one of which received counseling, was significantly in favor of the counseled groups, at better than the 1% level.

3. When non-counseled and counseled groups were compared, the counseled students increased their grade point average by an amount more than the non-counseled students with a significance at the 2% level.

Received October 1, 1948.

References

1. Beaumont, H. The evaluation of academic counseling. *J. higher Educ.*, 1939, 10, 79–82, 116.
2. Blackwell, E. B. An evaluation of the immediate effectiveness of the Testing and Guidance Bureau of the University of Texas. *J. educ. Res.*, 1946, 40, 302, 308.
3. Cowley, W. H. An experiment in freshman counseling. *J. higher Educ.*, 1933, 4, 245–248.
4. Freeman, H. J., and Jones, L. Final report of the long time effect of counseling low percentile freshmen. *Sch. & Soc.*, 1933, 38, 382–384.
5. Holladay, P. W. The long time effect of freshman counseling. *Sch. & Soc.*, 1929, 29, 234–236.
6. Lemon, A. C. *An experimental study of guidance and placement of freshmen in the lowest decile of the Iowa Qualifying Examination, 1925.* University of Iowa Studies in Educ. III (1927), 8, University of Iowa.
7. Newland, T. E., and Ackley, W. E. An experimental study of the effect of educational guidance on a selected group of high school sophomores. *J. exp. Educ.*, 1930, 5, 23–25.
8. Ward, J. R., and Tyler, L. E. A preliminary report of an evaluation of the Veterans Administration counseling service in the University of Oregon. *Amer. Psychol.*, 1947, 2, 416.
9. Williamson, E. G. The role of faculty counseling in scholastic motivation. *J. appl. Psychol.*, 1936, 20, 314–324.
10. Williamson, E. G. *A summary of studies in the evaluation of guidance.* Rep. Fifteenth Annl. Mtg. Coll. Personnel Assn., 1938, 73–77.
11. Williamson, E. G., and Bordin, E. S. Evaluating counseling by means of a control group experiment. *Sch. & Soc.*, 1940, 52, 434–440.
12. Williamson, E. G., and Bordin, E. S. The evaluation of vocational and educational counseling: a critique of the methodology of experiments. *Educ. & Psychol. Msmt.*, 1941, 1, 25–34.
13. Williamson, E. G., and Bordin, E. S. A statistical evaluation of clinical counseling. *Educ. & Psychol. Msmt.*, 1941, 1, 117–132.
14. Wrenn, C. G. *Recent research in counseling.* Rep. Sixteenth Annl. Mtg. Coll. Personnel Association, 1939, 88–94.

A Follow-up Study of Social Guidance at the College Level *

Margaret Glockler Aldrich

University of Missouri

In 1940, the author published a research report entitled "An Exploratory Study of Social Guidance at the College Level."[1] Early in 1948 it was decided to check the available records of the girls who as college freshmen (1939–1940) were the subjects for the experiment. It was felt that eight years would be sufficient for them to have completed their undergraduate careers. In checking the records it was found that only one girl was still in residence at the University of Minnesota in 1947–1948. She returned to school under the G. I. Bill and her transcript looks as if she might soon fulfill the requirements for graduation.

The original study was an attempt to compare two groups of freshmen girls who were all cases at the University Testing Bureau.[2] All of the girls went through the usual testing and counseling procedures of the Bureau. The experimental group received additional guidance in the social adjustment area and were directed toward participation in extracurricular activities. It consisted of at least one added interview with each girl in the experimental group stressing her social and activity life. In most cases this resulted in a definite contact with one or more of the activities in which the girl expressed an interest. The organizations had been contacted concerning the general need for good cooperation between various campus agencies. They did not know, however, that these girls were in any way "special cases." It seems safe to assume that the girls in the experimental group were also exposed to the usual social and extracurricular program in the same way that all freshmen girls are exposed.

[1] *Educational and Psychological Measurement.* Vol. II, No. 2, April, 1942, pp. 209–216.

[2] UTB is now called Student Counseling Bureau.

The control group had no added counseling but, of course, the girls in the group were free to make use of the University social and extra-curricular program. At the end of the school year both groups were retested on several personality scales and given a questionnaire. The conclusion reached at that time was: "All of these findings combine to indicate that, from this small sample, social guidance and directed participation in extra-curricular activities improve the 'social adjustment' of Freshmen girls as measured by personality scales and a questionnaire. Not only do the girls in the experimental group make greater mean gains, but they feel that they have more friends, participate in more activities, and are less critical of the social program than the control group. A treatment that makes people feel better satisfied with their social life is certainly worthy of further consideration." [3]

It should be pointed out that the study involved a very small sample, 31 experimental and 28 control subjects. Also, both groups were originally selected from the lower end of the distributions for freshmen girls on the Minnesota Inventory of Social Attitudes—Forms P and B and group activities in high school. They did not differ significantly, however, from the rest of the freshmen girls in mean ACE Psychological Examination score or in mean Cooperative English Test score. The experimental and control group were remarkably alike at the original testing on six objective measures (ACE, Coop. Eng., Social Beh., Social Pref., Rundquist-Sletto Inferiority Scale, and Bell Adjustment Inventory—social) and on high school group and individual activities. The control group was somewhat higher in high school scholarship rank.

Since the study covered only a brief period of time (9 to 12 months), it seemed worth while to re-study the groups after a period of eight years. This re-evaluation would indicate whether or not the gains revealed in the original study were ephemeral or were permanent.

The follow-up was confined to a check of the records kept by various campus agencies. The following agencies were contacted: Student Counseling Bureau; Student Activities Bureau; Bureau of Admissions and Records; Disciplinary Committee; Mental Hygiene Clinic of the Student's Health Service; and the Alumni Association.

In making the follow-up study, a new card was made for each girl with no indication of whether the girl belonged to the experimental or control group. All lists were sent to the agencies undesignated. This is of importance since several of the recordings involve judgments. When all of the data were collected the experimental and control groups were separated for analysis.

[3] *Op. cit.*, p. 216.

Results

Student Counseling Bureau Records. The Bureau records consist of a folder for each girl with her test results and a record dictated by the counselors of all counseling contacts in the Bureau. Table 1 summarizes the quantitative information and indicates that the experimental group had made, on the average, slightly more contacts over a slightly longer period of time.[4] The mean number of counseling contacts for both groups is considerably higher than the Bureau average of about two for these years.

Table 1

Student Counseling Bureau Contacts

Group	Mean No. of Contacts	Mean No. of Contacts After Retesting	Mean Duration of Contacts in Mos.
Control N = 24*	4.58	1.33	14.1
Experimental N = 31	5.74	1.65	15.5

* Four of the 28 girls in the control group were counseled by a counselor for the College of Science, Literature, and Arts. The folder of test results was kept by the U. T. B., but the interview records were kept in the S. L. A. office. Since these records are destroyed after five years, these girls had to be omitted from this part of the study.

Student Activities Bureau. In the years 1936–1946 the Student Activities Bureau kept records of the extra-curricular activities of all students in the University. The records were tabulated each quarter by the Bureau staff from their membership, committee, and officer lists and from publicity in the college newspaper. The director of the Bureau feels that the records are not too accurate and err in the direction of omitting activities.

The information from these cards has been summarized in Table 2 as mean number of activities, committees, and offices per year for the number of years the particular girl was in school. It must be emphasized that these are approximations and if anything underestimates. Nevertheless the results indicate that the girls in the experimental group participated in more activities, served on more committees, and held more offices than those in the control group.

[4] Statistical tests of significance of differences have not been computed because of the small N's and a belief that the chief value of the original study, and the present follow-up study, is to be found in the control group method of investigating the area of "social guidance."

Table 2

Student Activity Bureau Record

Group	Mean No. of Activities Per Year	Mean No. of Committees Per Year	Mean No. of Offices Per Year
Control N = 26*	.62	.03	.08
Experimental N = 30*	.96	.26	.27

* There were no cards in the files for two girls in the control group and one in the experimental group.

Gopher Record. A second source of activity record is the yearbook of the senior class, the *Gopher*. Each senior records his own activities for his years in college. The results from this source are very incomplete. It is only available for the girls who actually graduated and many of them did not have their picture and activity record included in the *Gopher*.

Table 3

Activity Record from Gopher (College Yearbook)

Group	Mean No. of Activities Listed	Mean No. of Committees Listed	Mean No. of Offices Listed
Control N = 8	2.5	.25	.38
Experimental N = 10	4.6	.60	1.00

Table 3 is based on the records of 18 girls who were included in the *Gopher* (a little over 50 per cent of those who graduated). The years '41, '42, '43, '44, and '46 were checked since these are the years listed for the graduates on the official transcripts. This rather skimpy evidence again points in the direction of greater activity for the experimental group. These data can be compared with the Activity Bureau record, Table 2, by dividing each mean by 4 to get the mean per year. The results are strikingly similar as shown in Table 4.

These results raise an interesting question which might be investigated further. There is a common idea that students tend to overestimate their activity record for publication. This small sample did not do this, particularly if we remember that there is some evidence that the SAB activity record is an underestimate.

It should be added that the records of Mortar Board (Senior Women's Honorary) were also checked for these years. Mortar Board picks its members from the entire junior class on the basis of scholarship, leadership, and service. Through the years at Minnesota this group has tended to include the leaders in extra-curricular activities if their grades were up to a certain fixed level. Two girls from this study were elected to the 1943 chapter of Mortar Board. They were both members of the experimental group.

Table 4

SAB Activity and Gopher Record Compared

Group	Mean No. of Activities Per Year		Mean No. of Committees Per Year		Mean No. of Offices Per Year	
	Act.	Goph.	Act.	Goph.	Act.	Goph.
Control	.62	.83	.03	.08	.08	.09
Experimental	.96	1.15	.26	.15	.27	.25

Bureau of Admissions and Records. Data from the Bureau of Admissions and Records consisted of a transcript for each girl. Table 5 summarizes these data.

The academic records of the two groups are similar although the control group had a somewhat higher average. It might be well to recall that they also had a slightly better high school academic record.

Table 5

Information from Official Transcript

Group	Per Cent Graduated	Mean H.P.R.*	Mean No. of Quarters at Minn.
Control N = 15	54	1.51	8.86
Experimental N = 18**	58	1.12	8.87

* H.P.R. = honor point ratio = honor points/credits, where for each credit of A, 3, B, 2, C, 1, D, 0, and F, −1 honor points are given. These were calculated only from University of Minnesota grades and for undergraduate work.

** This figure omits two A.A. (Associate of Arts) degrees: a two year degree granted by the General College. There were 5 girls who did some work in General College. Their records are not included in Column 2 since General College grades are not directly comparable to grades in other colleges.

Disciplinary Committee Records. The list of girls was sent to the head of the Disciplinary Committee of the university. He reported that none of the 59 names was recorded in the files of that committee.

Mental Hygiene Clinic. The list of names was also sent to the head of the Mental Hygiene Clinic in the Students' Health Service. He had the names checked against the clinic records. Six girls had contacted the Clinic, three from the Control Group and three from the Experimental. In each case he made an estimate of severity of diagnosis with the result that those from the Control Group were labelled "severe" whereas none from the Experimental Group were so designated.

Although the psychiatrist reports that about 5 per cent of the University population would like to make contact with the Clinic, he estimates that through the years the Clinic has had facilities for only about 3 per cent. This is much lower than the 10 per cent of both the experimental and control group who went to the Clinic. This might be explained by the original selection of the groups from the lower end of the distributions on the personality scales. One might hypothesize that the social guidance did little to prevent the development of problems requiring mental hygiene but that these problems were less severe for the girls who had the earlier specialized help. Obviously this is little more than a hunch.

Table 6

Per Cent of Married Graduates *

	N	Married	N	Not Married
Control N = 14	10	71%	4	29%
Experimental N = 18	11	61%	7	39%

* Includes A.A. degrees. Total N = 32. The alumni office records, however, did not list as graduates three girls whose transcripts indicate that they did receive degrees.

Alumni Association Records. The records of the Minnesota Alumni Association are kept only for the students who actually graduate from the University of Minnesota. For each graduate there is a fairly complete record of address and married name. At the time of this study the latter record is summarized in Table 6. Clearly from these incomplete records a higher percentage of the control group married. If we consider marriage an indication of social adjustment, the control group (at least those who graduated) is better adjusted. This is the only scrap of evidence in favor of the control group.

Summary

This follow-up of social guidance can be summarized in three sections.

1. Those who received special guidance with social problems exceeded the control group in: (a) the number of contacts with the Student Counseling Bureau; (b) the mean number of college activities, committees, and offices; (c) the percentage graduating from the University of Minnesota; and (d) a less severe diagnosis for those who contacted the Mental Hygiene Clinic.

2. The groups were much alike in: (a) the mean number of months over which the contacts with the Student Counseling Bureau were made; (b) the number of quarters in residence at the University of Minnesota; and (c) the number of girls who contacted the Mental Hygiene Clinic.

3. The control group was slightly higher than the experimental group in: (a) mean honor point ratio; and (b) the percentage of the graduates listed in the Alumni Bureau files as married.

The small numbers in both groups make more detailed statistical analysis of questionable value. From the data available, however, there is an indication that the gains originally reported for the socially guided group continued throughout their college residence. Again, the tentative conclusion of the original study can be re-emphasized with the caution mentioned in the last sentence of that study, "the problem was, however, essentially an investigation of a method and as such the results should be emphasized only as a justification for the further use of the method."

Memory in Radio News Listening

Thomas W. Harrell, Donald E. Brown, and Wilbur Schramm

University of Illinois

Questions of practical importance have arisen in the field of radio involving the extent to which a listener is able to remember what he hears on a newscast. The newscaster is anxious to know how tightly he can "pack" his newscast—how many stories he can put into a given time without giving his audience more than they can absorb. Beyond that, he wants to know the effect on memory of repetition within the newscast. He is interested in what kinds of subject matter and what treatments of those are remembered better than others. He would like to know whether his audience listens for "index words," whether it remembers names and details, whether it remembers items far removed in locale as well as it remembers items originating nearby. Finally, he would like to know, if possible, what kinds of items discriminate least between good memories and poor memories, and therefore, so far as the factor of memory is concerned, are mass materials for a mass medium.

In a situation wherein the average adult American listens more than three hours a day to the radio, between 10 and 25 per cent of this time to radio news, these questions become of social as well as professional importance. The study reported here was undertaken in an attempt to provide experimental data in an area where the hunch and the thumb have ruled.

Method

Two entirely different news broadcasts each containing 20 stories were written by an experienced news editor. These were designated as broadcasts IA and IIA. These 20 stories were reduced in size but with care being taken not to omit any important detail, to permit the addition of 10 more stories making a total of 30 stories in newscasts IB and IIB. (All newscasts actually ran $12\frac{1}{2}$ minutes in order to make them the same length as are most commercial casts on the radio. It was not thought necessary for the purpose of this study to insert commercials.)

For the next series of newscasts which were written in a highly compressed style each of these 30 stories was further reduced in length which provided time for the addition of 10 new stories, making a total of 40 (newscasts IC and IIC).

All six newscasts were transcribed. An experienced announcer read the casts, transcribing only two per day to avoid staleness. Each cast was transcribed to a platter, tape, and wire. It proved more convenient to use the tape transcriptions except for one presentation of a platter recording.

The casts were fictitious but plausible. Real happenings were not presented because there would then have been some persons who were already more familiar than others with the event. That the casts were highly realistic was suggested by the questions of some subjects, who, even though assured of the contrary, would inquire whether "there was anything to" one or more of the stories.

Memory tests were constructed for each cast. Four alternative multiple choice questions with single best answer were used. The scoring formula, Right − ⅓ Wrongs, was used to correct for chance successes. One question was asked on each story so there were 40 questions each on casts IC and IIC, 30 on casts IB and IIB, and 20 on casts IA and IIA. All the questions on A casts were repeated in B casts, and all questions in B casts were repeated in C casts. The aim was to make the questions central to the story and as easy as possible while at the same time assuring discrimination from guessing.

Two casts each were presented to ten groups of subjects. Each I cast was presented to a group which also heard a II cast of a different number of stories. The order of presentation was reversed from one session to another because of the possibility of a practice or fatigue effect.

The method is recognized as being not true to life. In the first place the casts were fictitious. In the second place the subjects were assembled and had fewer distractions than do radio listeners ordinarily. It is expected that the experimental conditions would yield a maximum of what could be remembered in real life. It is believed that when listening to news on the radio the majority of listeners do not give as good attention as in the experimental setting. On the other hand, one could conjecture that because of the fictitious nature of the news it would not be attended quite so well as if it were real, so there would be some compensating effect to the extraordinary attention. Some thought was given to using real news casts and actual listeners, but the expense of such a study was found to be prohibitive.

Each group of listeners was told the purpose of the study, that there would be a memory test after each cast, and that there would be a preference question after both casts.

An effort was made to choose as subjects adults similar in education to the average of the American radio listening audience. The majority of subjects were enlisted men and women of the United States Air Force.

These Air Force enlisted man and women were members of the permanent party of a base. Their average educational level was in the neighborhood of 10th grade. Their standard scores on the Army General Classification Test ranged from approximately 90–115, which is practically the range of the complete adult population of military age.

The subjects also included two groups of nonacademic employees and three groups of students at the University of Illinois. One of the groups of nonacademic employees was a group of groundsmen whose educational level was similar to that of the Air Force subjects. The second group of nonacademic subjects were supervisors whose educational level ranged from high school graduation to college graduation. The student subjects were undergraduates. The subjects were somewhat above the average of the American public in education and consequently above the average of the radio listening audience, but since over half of the subjects were within the average range in education, they were regarded as satisfactory for the experiment.

Results: I. Memory and the Number of Stories

A reasonable hypothesis is that if a listener is presented a progressively increasing number of items within a fixed period of time, he will remember a progressively smaller proportion of them. The results bear this out, as Table 1 shows.

Table 1

Memory for Broadcasts

Test	N	Mean %	σ_{mean}	σ_{dis} %	Mean Raw Score	σ Raw Score
A (19–20 items)*	320	54.5	1.20	21.55	10.9	4.31
B (29–30)*	264	49.3	1.12	18.30	14.8	5.49
C (40)	308	45.9	1.09	18.50	18.3	7.40

* One item had to be omitted from scoring in two sets of the tests.

Table 2 shows the statistical significance of these differences.

A further test of these figures is given in Table 3, which shows positive and significant correlations between each pair of test scores.[1] This indicates that listeners who were high on one test tended also to be high on the other. Therefore, the listeners must have been attending, and the tests were measuring the same thing, whatever it was they were measuring.

[1] The variations in size of the coefficients do not make sense to the investigators, and are presumably due to chance.

It seems to be a tenable hypothesis, then, that a listener remembers a smaller proportion of items in a fixed-time newscast if the number of items is increased from 20 to 30 to 40. The question then follows: where is the point of insufficient return? Where does the memory curve fall off so sharply that the newscaster may conclude he has overpacked his newscast?

Table 2
Probability that Memory Differences May Be Due to Chance

Differences	Probability
A and B (19–20 and 30)	.0012
B and C (29–30 and 40)	.013
A and C (19–20 and 40)	.0000

Table 3
Coefficients of Correlation Between Scores on Each Pair of Tests

Tests	r	N
IA–IIB	.63	77
IA–IIC	.41	127
IIA–IB	.72	61
IIA–IC	.43	55
IB–IIC	.51	80
IIB–IC	.68	46

While the differences shown in Table 1 are significant, they are nevertheless slight. In fact, they are so slight that a listener actually remembers more items from a 30-item cast than from a 20-item cast, from a 40 than from a 30. In a 20 item newscast 11 stories are remembered, in a 30 story newscast 15 stories are remembered, and 18 stories are remembered in a 40 story newscast. It must be concluded therefore, that there is nothing in this evidence, so far as the factor of memory goes, to lead a newscaster to set an arbitrary limit below 40 items in a 12½ minute newscast if his material justifies that many items. The factor of audience preference, however, bears strongly on this point, as the next section of this report will show.

For what it is worth, these figures suggest that a listener remembers a few minutes after a newscast has been heard, about half the items in the newscast. This suggests several related questions, such as the kinds of material that are remembered best, the effect of repetition, and the kinds of cues that arouse the best learning response in news listening

situation. These questions are discussed in sections III, IV, and V of this report.

Results: II.   Preference and the Number of Stories

A preference question was asked at the end of each pair of casts. The results are shown in Table 4.

Table 4

Preferences for Broadcasts

	N	%	Probability that True % = 50
A (20)	49	52	
B (30)	45	48	.3483
A (20)	135	74	
C (40)	48	26	.0000
B (30)	83	66	
C (40)	43	34	.0001

These figures indicate that the broadcast with 40 stories was clearly liked less than those of 20 and 30 stories.   Approximately three out of four people preferred the 20-item casts to the 40-item casts.   Almost exactly two out of three persons preferred the 30-item casts to the 40-item casts.   The slight preference for 20 items as compared with 30 is statistically insignificant.   These figures suggest that the memory effort involved in listening attentively to a 40-item newscast, though quite possible for the average listener, is not popular; and this provides good reason for the newscaster to limit his number of items to 30, perhaps still better to 20.

Results: III.   Memory and Repetition of News Facts

Sixteen questions were so designed as to repeat facts, to be tested, oftener in one cast than in another.   When test results on these questions are compared, there is no significant trend discernible, as Table 5 shows.

On the basis of these results, the hypothesis can be advanced that repetition of facts in a newscast has no significant effect on audience memory of those facts.   It may be, of course, that uncontrolled variables entered into this result.   The number of stories in newscast may have some influence on the effectiveness of repetition.   It would seem, however, that whatever influence is present here might work in the direction of making repetition appear to be more important than it really is.   This is true because there is more repetition in the casts with fewer stories.

Table 5

Per Cent of Listeners Answering Question Correctly, Compared with
Number of Times Answer Was Repeated in Cast

Cast IA Times Rep.	%	Cast IB Times Rep.	%	Cast IC Times Rep.	%
3	91	3	82	2	80
2	78	2	54	1	76
2	60	1	54	1	59
4	51	3	52	3	48
2	38	1	59	1	62
2	84	2	64	1	53
3	81	2	79	2	77
3	78	2	68	1	47

Cast IIA		Cast IIB		Cast IIC	
3	86	2	90	2	80
2	85	1	61	1	64
2	60	1	62	1	54
3	66	1	68	1	49
2	60	2	58	1	35
2	56	1	57	1	55
3	56	2	44	1	35
3	91	1	80	1	97

It has been shown that there is a slight tendency for memory to be better
for any single story in a cast with 20 stories as compared to the cast with
40 stories. Since in spite of this the statistical results show repetition
to be of slight if any importance there is all the more reason to doubt the
effectiveness of this kind of repetition. It must be remembered, of
course, that this was not overt or enforced repetition; it was not done in
the jangling fashion of "LS/MFT" or even in the style, "I'll repeat that
name again." Those signposts may make repetition more effective in
creating a response that leads to memory. Furthermore, it may be that
repetition is more effective in other repetitive situations—for example,
when a story is heard on more than one newscast.

Results: IV. Memory and Subject Matter

The questions were divided by subject matter, and test scores com-
pared on that basis. This is not an artificial division, inasmuch as most
newscasts are compartmentalized by some kind of subject matter dis-
tinctions. The results are shown in Table 6.

Because of the small number of human interest questions, the difference between that score and others is not statistically significant. Between the mean per cents right on spectacular events and public affairs, the difference is significant at the 5% level; between public affairs and name items, at the 1% level.[2] It appears, then, that name items are hardest to remember; that public affairs items are harder to remember than stories of fires, windstorms, wrecks, murders, lynchings, and other spectacular events; and further tests may show that human interest items are easiest of all to remember.

Table 6
Average Scores on Questions Classified by Content

	No. of Stories	Mean %	σ_{mean} %'s
Human Interest	11	85	3.21
Spectacular Events	46	72	2.42
Public Affairs	36	63	2.56
Name Items	79	53	1.94

Results: V.   Memory and Mass Audiences

One of the sets of test was analyzed on the basis of how well each question discriminated between persons who did well on their two tests, and therefore may be supposed to have good memories, and persons who did poorly on their two tests, and therefore may be supposed to have less good memories.   In order to do this, each question was ranked according to the difference between the number of participants below median and the number above median.   When this was done, the middle half of the questions was discarded and attention focussed on the highest and lowest quartiles.   It was assumed that the top quartile contains questions which most clearly show the difference between the best and the poorest memories; therefore, that the material being tested in this quartile is material which is more difficult and less well adapted to mass audiences. It was assumed also that the lowest quartile contains questions which least clearly show the difference between best and poorest memories; and therefore, that the material being tested in this quartile is least difficult and best adapted to mass audiences.   The material in the two quartiles was then analyzed both in terms of subject matter and of the approach to that subject matter used in framing that question.   Table 7 gives part of that analysis.

[2] Spectacular events and public affairs: t = 2.51, with 80 degrees of freedom.   Public affairs and name items: t = 2.99 with 113 degrees of freedom.

On the basis of this analysis and further examination of items and questions, several hypotheses may be set forth.

For one thing, public affairs appears to be the subject matter which chiefly discriminates between listeners who have good memories and listeners who do not; whereas materials involving crime, disaster, and human interest are remembered almost as well by poor memories as by good ones.

Table 7

Analysis of Items Which Proved to be Most and Least Discriminatory Between Good and Poor Memories

Subject Matter	Kind of Information Required by Question	Locale	Cues
(Highest quartile—most discriminatory)			
Public affairs	Details of political action	Foreign	Russia
Public affairs	Details of violent action	Foreign	Palestine
Public affairs	Names plus details of political action	Foreign	Inter-American affairs
Public affairs	Details of economic policy	National	Taxes
Public affairs	Names of political classifications	National	Politics
Disaster	Details of accident	National	Airplane
Disaster	Details of accident	Regional	Mayor of nearby city
Public affairs	Details of political action	Regional	Methodist minister
Human interest	Names and cities	Regional	State bar association
Human interest	Name of war	Regional	Civil War
(Lowest quartile—least discriminatory)			
Public affairs	Details of quotation	National	Forrestal—Alaska-Russia
Disaster	Details of cause of fire	National	Children—fire
Disaster	Details of cause of accident	Regional	Old man—teacher
Public affairs	Details of cause of strike	Regional	Strike—name of nearby town
Crime	Details of violent action	Regional	Negro—lynching—murder
Crime	Name of town	Regional	Escaped convict
Human interest	Name of person	Regional	State farmers union
Human interest	Details of prize won	Regional	Hollywood—Cinderella story
Human interest	Details of divorce action	National	Hollywood-name of well-known star
Crime	Details, nature of crime	National	American sailors assault

There is a slight indication that names may discriminate more than details, but the essential difference seems rather to be the *kind* of detail. Political detail seems to be more discriminatory than sensational detail. It may well be that such a combination as foreign names and political details, as in one of the upper quartile questions, may put the greatest challenge to listener's memories.

It appears that events far removed in locale are more likely to discriminate between good and poor memories than events near at hand. It will be noticed that there are no foreign stories in the lowest quartile, and that one of the stories there classified as "national" is about Hollywood, a locale which mass communications have brought next door to all America.

One theory of radio news listening is that the listener puts into effect his own selective mechanism to parallel the newspaper reader's use of headlines or the magazine reader's use of the table of contents. That is, it is conjectured that the radio audience listens at a rather low level of attentiveness until he hears an "index" word or phrase which triggers a response, raises the level of attention, and causes perception to take place. With this theory in mind, it is interesting to look at the column of "cues" in Table VII. These are the words which seem to "stick out" from the stories, the ones which might serve as index words or cues to create a response in case that process is the one in effect. It will be noticed that the stories in the lowest quartile have a high incidence of rather sensational or familiar cues—Hollywood, Danny Kaye, children burning, old age, strike, lynching, murder, escaped convict, towns nearby. The stories in the highest quartile, on the other hand, have rather more sophisticated cue words—the Inter-American situation, politics, taxes, Palestine.

As far as the memory factor goes, then, it would seem to be possible to hypothesize a formula for a newcaster's approach to the lowest common denominator, and therefore to a mass audience. That formula would be about the same as the one used by many newspapers which have reached mass circulations—sensation, crime, disaster, human interest; public affairs subordinated or treated in a sensational manner; an emphasis on nearby places and familiar names; and a plentiful sprinkling of interest-attracting words, names, and phrases.

A word of caution may be unnecessary here. Nevertheless, the investigators wish to make it clear that they do not consider these facts to be reason for subordinating public affairs in newscasts, or for sensationalizing and infantizing all copy on the grounds that such is the least common denominator of the mass audience and radio is a mass medium. That is no more required of a newscast than it is required of all news-

papers. Nor is it the import of these results. Rather, these results point to further study of the use of public affairs news on the air—how it may be made useful and effective for the part of the audience which needs it, without loss of either truth or dignity; the extent and connection to which names and details can be used when important for the audience's information; and the boundaries, if any, between kinds of material which can best be presented to the ear or to the eye. Questions like these will yield to experimental approach, and radio will grow in its public service if the results of such experiments can be incorporated into practice.

Summary

1. An audience remembers a proportionately smaller percentage of the items in a 15-minute newscast as the number of items is increased from 20 to 30 to 40. This difference, however, is slight—so slight that actually more items are remembered from the 30-item newscast than from the 20, more from the 40 than from the 30.

2. An audience has a decided preference, however, for newscast with 20 or 30 items over one with 40 items.

3. Repetition of facts within a newscast has not been shown to have a significant effect on audience memory.

4. Human interest and spectacular stories of crime and disaster are remembered better than are stories of public affairs.

5. Insofar as the factor tested is concerned, the appeal to a mass audience by radio news is similar to the appeal of certain sensational newspapers which have reached mass audiences. Results of this study indicate that human interest and spectacular events are remembered by the mass audience, whereas such serious subject matter as public affairs is remembered less well by the part of the population which is not gifted with good memories. Nearby events are more likely to be remembered by the mass audience than events of distant origin. Details and names do not make for mass remembrance, and details of political events and foreign names in a public affairs story are especially hard to remember. "Index words" of a sensational or familiar nature are also helpful in penetrating the memories of the mass audience.

Tables for Use with the Flesch Readability Formulas

James N. Farr and James J. Jenkins

University of Minnesota

Increased emphasis is being given to measurements of the readability of communications in many fields. A promising approach which is being widely used and studied is that set forth by Flesch [1,2] which involves the use of syllable counts, sentence lengths, percentage of personal words and percentage of personal sentences to yield two indexes. One index is "Reading Ease" or level of difficulty and the other is "Human Interest."

In order to facilitate the use of these formulas, the writers have tabled the values for them. The tables are simple to use. Table 1, "Reading Ease," is entered vertically by average sentence length and horizontally by the number of syllables per one hundred words. The index figure is given at the point of intersection of the row and column entries. For example, if a sample of one hundred words contains 133 syllables and has an average sentence length of 25 words, the "Reading Ease" index equals 69. This index number may then be interpreted directly in terms of difficulty by Flesch's table.[2]

In like manner the "Human Interest" table (Table 2) is entered vertically by percentage of personal sentences and horizontally by the percentage of personal words to obtain that index. For example, if a sample has thirteen personal words per one hundred words and ten percent of the sentences are personal, the "Human Interest" index is equal to 50. This figure may be directly interpreted in terms of interest by Flesch's table.[2]

Several checks were made to insure the accuracy of the tables. The outer edge indexes were computed separately by the writers. One writer obtained the tabled values by use of the subtractive constant for columns; the other used the subtractive constant for rows. Both writers checked the work by use of the subtractive constant for selected diagonals.

Since the formulas are both straight-line functions, simple abacs may be easily constructed for use in situations where only approximations are needed.

[1] Flesch, R. *The art of plain talk.* New York: Harper and Brothers, 1946.

[2] Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221–233.

James N. Farr and James J. Jenkins

Table 1

Flesch Reading Ease Index Table

Average Sentence Length	Syllable Count per Hundred Words																													
	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149
9	96	95	94	94	93	92	91	90	89	89	88	87	86	85	84	83	83	82	81	80	79	78	78	77	76	75	74	73	72	72
10	95	94	93	93	92	91	90	89	88	88	87	86	85	84	83	82	82	81	80	79	78	77	77	76	75	74	73	72	71	71
11	94	93	92	92	91	90	89	88	87	87	86	85	84	83	82	81	81	80	79	78	77	76	76	75	74	73	72	71	70	70
12	93	92	91	91	90	89	88	87	86	86	85	84	83	82	81	80	80	79	78	77	76	75	75	74	73	72	71	70	69	69
13	92	91	90	90	89	88	87	86	85	85	84	83	82	81	80	79	79	78	77	76	75	74	74	73	72	71	70	69	68	68
14	91	90	89	89	88	87	86	85	84	83	83	82	81	80	79	78	78	77	76	75	74	73	72	72	71	70	69	68	67	67
15	90	89	88	88	87	86	85	84	83	82	82	81	80	79	78	77	77	76	75	74	73	72	71	71	70	69	68	67	66	66
16	89	88	87	87	86	85	84	83	82	81	81	80	79	78	77	76	76	75	74	73	72	71	70	70	69	68	67	66	65	65
17	88	87	86	86	85	84	83	82	81	80	80	79	78	77	76	75	75	74	73	72	71	70	69	69	68	67	66	65	64	64
18	87	86	85	85	84	83	82	81	80	79	79	78	77	76	75	74	74	73	72	71	70	69	68	68	67	66	65	64	63	63
19	86	85	84	83	83	82	81	80	79	78	78	77	76	75	74	73	72	72	71	70	69	68	67	67	66	65	64	63	62	61
20	85	84	83	82	82	81	80	79	78	77	77	76	75	74	73	72	71	71	70	69	68	67	66	66	65	64	63	62	61	60
21	84	83	82	81	81	80	79	78	77	76	76	75	74	73	72	71	70	70	69	68	67	66	65	65	64	63	62	61	60	59
22	83	82	81	80	80	79	78	77	76	75	75	74	73	72	71	70	69	69	68	67	66	65	64	64	63	62	61	60	59	58
23	82	81	80	79	79	78	77	76	75	74	74	73	72	71	70	69	68	68	67	66	65	64	63	63	62	61	60	59	58	57
24	81	80	79	78	78	77	76	75	74	73	72	72	71	70	69	68	67	67	66	65	64	63	62	61	61	60	59	58	57	56
25	80	79	78	77	77	76	75	74	73	72	71	71	70	69	68	67	66	66	65	64	63	62	61	60	60	59	58	57	56	55
26	79	78	77	76	76	75	74	73	72	71	70	70	69	68	67	66	65	65	64	63	62	61	60	59	59	58	57	56	55	54
27	78	77	76	75	75	74	73	72	71	70	69	69	68	67	66	65	64	64	63	62	61	60	59	58	58	57	56	55	54	53
28	77	76	75	74	74	73	72	71	70	69	68	68	67	66	65	64	63	63	62	61	60	59	58	57	57	56	55	54	53	52
29	76	75	74	73	72	72	71	70	69	68	67	67	66	65	64	63	62	61	61	60	59	58	57	56	56	55	54	53	52	51
30	75	74	73	72	71	71	70	69	68	67	66	66	65	64	63	62	61	60	60	59	58	57	56	55	55	54	53	52	51	50
31	74	73	72	71	70	70	69	68	67	66	65	65	64	63	62	61	60	59	59	58	57	56	55	54	54	53	52	51	50	49
32	73	72	71	70	69	69	68	67	66	65	64	64	63	62	61	60	59	58	58	57	56	55	54	53	53	52	51	50	49	48
33	72	71	70	69	68	68	67	66	65	64	63	63	62	61	60	59	58	57	57	56	55	54	53	52	52	51	50	49	48	47
34	71	70	69	68	67	67	66	65	64	63	62	61	61	60	59	58	57	56	56	55	54	53	52	51	51	50	49	48	47	46
35	70	69	68	67	66	66	65	64	63	62	61	60	60	59	58	57	56	55	55	54	53	52	51	50	49	49	48	47	46	45
36	69	68	67	66	65	65	64	63	62	61	60	59	59	58	57	56	55	54	54	53	52	51	50	49	48	48	47	46	45	44
37	68	67	66	65	64	64	63	62	61	60	59	58	58	57	56	55	54	53	53	52	51	50	49	48	47	47	46	45	44	43
38	67	66	65	64	63	63	62	61	60	59	58	57	57	56	55	54	53	52	52	51	50	49	48	47	46	46	45	44	43	42

Table 1—Continued

Syllable Count per Hundred Words

Average Sentence Length	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179
9	71	70	69	68	67	67	66	65	64	63	62	61	61	60	59	58	57	56	56	55	54	53	53	51	50	50	49	48	47	46
10	70	69	68	67	66	66	65	65	64	63	62	61	60	59	58	57	56	55	55	54	53	52	52	50	49	49	48	47	46	45
11	70	68	68	67	66	65	64	63	63	62	61	60	59	58	57	56	55	55	54	53	52	51	51	49	48	48	47	46	45	44
12	69	67	67	66	65	64	63	62	62	61	60	59	58	57	56	55	54	54	53	52	51	50	50	48	47	47	46	45	44	43
13	68	66	66	65	64	63	63	62	61	60	59	58	57	56	55	54	53	53	52	51	50	49	49	47	46	46	45	44	43	43
14	66	65	64	64	63	62	62	61	60	59	58	57	56	55	54	53	52	52	51	50	49	48	48	46	46	45	44	43	42	42
15	66	65	64	63	62	61	61	60	59	58	57	56	55	54	53	52	51	51	50	49	48	47	47	46	45	44	43	43	41	41
16	65	65	63	63	62	61	60	59	58	57	56	55	54	53	52	51	51	50	49	48	47	46	46	44	43	43	42	42	40	40
17	63	63	61	60	59	58	57	57	56	55	54	53	52	51	50	50	49	48	47	46	46	45	44	43	42	42	41	41	39	39
18	62	61	60	59	58	57	57	56	55	54	53	52	51	50	50	49	48	47	46	45	45	44	43	42	41	41	40	40	38	38
19	61	60	59	58	57	56	56	55	54	53	52	51	50	49	48	47	46	46	45	44	43	42	42	41	40	39	39	38	37	36
20	60	59	58	57	57	56	55	54	53	52	51	50	49	48	48	47	46	45	44	44	43	42	41	40	39	38	38	37	36	35
21	59	58	57	56	55	54	54	53	52	51	50	49	48	47	46	45	44	43	43	42	41	41	40	39	38	37	37	36	36	34
22	58	57	56	55	54	53	53	52	51	50	49	48	47	47	46	45	44	43	42	41	40	40	39	38	37	36	35	35	34	33
23	57	56	55	54	54	53	52	51	50	49	48	47	47	46	45	44	43	42	41	40	39	39	38	37	36	35	34	34	33	32
24	56	55	54	53	53	52	51	50	49	48	47	46	46	45	44	43	42	41	40	39	38	38	37	36	35	34	33	33	32	31
25	55	55	53	52	52	51	50	49	48	47	46	45	45	44	43	42	41	40	39	38	37	37	36	35	34	33	32	32	31	30
26	55	54	53	52	51	50	49	48	47	46	45	44	44	43	42	41	40	39	38	37	36	36	35	34	33	32	31	31	30	29
27	53	53	51	50	50	49	48	47	46	45	44	43	43	42	41	40	39	38	37	36	35	35	34	33	32	31	30	30	29	28
28	52	51	50	49	49	48	47	46	45	44	43	42	42	41	40	39	38	37	36	35	34	33	33	32	31	30	30	29	28	28
29	51	50	49	48	48	46	46	45	44	43	42	41	40	40	39	38	37	36	35	34	33	33	32	31	30	30	29	28	27	26
30	49	49	48	47	46	45	45	44	43	42	41	41	40	39	38	37	36	35	34	34	33	32	31	30	29	28	27	27	26	25
31	48	48	47	46	45	45	44	43	42	41	40	40	39	38	37	36	35	34	33	33	32	31	30	29	28	27	26	26	25	24
32	47	47	46	45	44	44	43	42	41	40	39	39	38	37	36	35	34	33	32	32	31	30	29	28	27	26	25	25	24	23
33	46	46	45	44	43	43	42	41	40	39	38	38	37	36	35	34	33	32	31	31	30	29	28	27	26	25	24	24	23	22
34	45	45	44	43	43	42	41	40	39	38	37	37	36	35	34	33	32	31	30	30	29	28	27	26	25	24	23	23	22	21
35	45	44	43	42	42	41	40	39	38	37	36	36	35	34	33	32	31	30	30	29	28	27	26	25	24	23	22	22	21	20
36	43	43	42	41	41	40	39	38	37	36	35	35	34	33	32	31	30	29	28	28	27	26	25	24	23	22	21	20	20	19
37	43	42	41	40	40	39	38	37	36	35	34	34	33	32	31	30	29	28	27	27	26	25	24	23	22	21	20	20	19	18
38	41	41	40	39	38	37	36	36	35	34	33	32	32	31	30	29	28	28	27	26	25	24	23	22	22	21	20	19	18	17

Table 2

Flesch Human Interest Index Table

								Percentage of Personal Words													
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0	00	04	07	11	15	18	23	25	29	33	30	40	44	47	61	55	58	62	65	69	73
2	01	04	08	12	15	19	22	26	30	33	37	41	44	48	52	55	59	62	66	70	73
4	01	05	09	12	16	19	23	27	30	34	38	41	45	49	52	56	59	63	67	70	74
6	02	06	09	13	16	20	24	27	31	35	38	42	40	49	53	56	60	64	67	71	75
8	03	06	10	13	17	21	24	28	32	35	39	42	46	50	53	57	61	64	68	72	75
10	03	07	10	14	18	21	25	29	32	36	39	43	47	50	54	58	61	65	69	72	76
12	04	07	11	15	18	22	26	29	33	36	40	44	47	51	55	58	02	86	69	73	76
14	04	08	12	15	19	23	26	30	33	37	41	44	48	52	55	59	63	66	70	73	77
16	05	09	12	16	20	23	27	30	34	38	41	45	49	52	56	00	03	67	70	74	78
18	06	09	13	17	20	24	27	31	35	38	42	46	49	53	57	60	64	67	71	75	78
20	06	10	14	17	21	24	28	32	35	39	43	46	50	54	57	61	64	68	72	75	79
22	07	11	14	18	21	25	29	32	36	40	43	47	51	54	58	61	65	69	72	76	80
24	08	11	15	18	22	20	29	33	37	40	44	48	51	55	58	62	66	69	73	77	80
26	08	12	15	19	23	26	30	34	37	41	45	48	52	55	59	63	66	70	74	77	81
28	09	12	16	20	23	27	31	34	38	42	45	49	52	56	60	63	67	71	74	78	81
30	09	13	17	20	24	28	31	35	39	42	40	49	53	57	60	64	68	71	75	78	82
32	10	14	17	21	25	28	32	35	39	43	46	50	54	57	61	65	08	72	75	79	83
34	11	14	18	22	25	29	32	36	40	43	47	51	54	58	62	65	69	72	76	80	83
36	11	15	19	22	26	29	33	37	40	44	48	51	55	59	62	66	69	73	77	80	84
38	12	16	19	23	26	30	34	37	41	45	48	52	56	59	63	66	70	74	77	81	85
40	13	16	20	23	27	31	34	38	42	45	49	53	56	60	63	67	71	74	78	82	85
42	13	17	20	24	28	31	35	39	42	40	50	53	57	60	04	68	71	75	79	82	86
44	14	17	21	25	28	32	30	39	43	47	50	54	57	61	65	08	72	76	79	83	87
46	14	18	22	25	29	33	36	40	44	47	51	54	58	62	85	09	73	76	80	84	87
48	15	19	22	26	30	33	37	41	44	48	51	55	59	62	66	70	73	77	81	84	88
50	16	19	23	27	30	34	38	41	45	48	52	56	59	63	07	70	74	77	81	85	88
60	19	23	26	30	33	37	41	44	48	52	55	59	62	66	70	73	77	81	84	88	92
70	22	26	29	33	37	40	44	47	51	55	58	62	66	69	73	77	80	84	87	91	95
80	25	29	32	36	40	43	47	51	54	58	62	65	69	72	76	80	83	87	91	94	98
90	28	32	36	39	43	46	50	54	57	61	65	68	72	76	79	83	86	90	94	97	X*
100	31	35	39	42	46	50	53	57	60	64	68	71	75	79	82	86	90	93	97	X*	X*

The left axis is labeled *Percentage of Personal Sentences*.

* X indicates 100 or over.

Inasmuch as these tables permit rapid and accurate determination of the Flesch index values and eliminate virtually all calculations previously involved, it is hoped that more research on the applicability and utility of the formulas will be undertaken.

Book Reviews

Bowler, Earl M., and Dawson, Frances Trigg. *Counseling employees.* New York: Prentice-Hall, 1948. Pp. xi+247. $4.00.

The authors state that this book is an answer to the well founded desire of employee counselors for a handbook written by practical people in down to earth style. Many readers will disagree.

Psychologists are not apt to be favorably impressed by a twenty-five degree merit rating scale, consideration of Cardall's Practical Judgment Test as a personality test, statements such as "It is not unusual for a good counselor to be called Mr. Anthony," and frequent use of generalities for which little if any experimental evidence is cited or available. Industrialists are not apt to agree that handicapped persons should be employed to prevent them from developing competitive companies, and that current job salaries are low.

This reviewer does not believe that publication of the book will improve the theory or practice of counseling.

C. E. Jurgensen

Minneapolis Gas Company

Kessler, Henry H., M.D. *Rehabilitation of the physically handicapped* New York: Columbia University Press, 1947. Pp. 251. $3.50.

Associated with the New Jersey Rehabilitation Commission from 1919 until 1941, at which time he entered the Navy to continue his rehabilitation activities, Dr. Henry H. Kessler has had a peculiar opportunity to participate in an integrated approach to the problem of seeing a person through from illness or injury to a job—in a word, rehabilitation. *Rehabilitation of the Physically Handicapped* is a general survey of the problems encountered in and the services that constitute an adequate rehabilitation program. The author presents his interpretation of the needs of the disabled and the many unsolved problems in rehabilitation as evidenced during twenty-eight years of active experience in this field. For a general treatment of vocational rehabilitation this publication has no equal.

The book is divided into four general sections. Part one describes the problems of the physically handicapped in general with special treatment of the crippled child, injured worker, disabled veteran and the chronic disabled. Social attitudes and legislation have in general crystallized around these groups of handicapped persons. After a critical review of the concept of physical fitness the author concludes that "physical

disability has no meaning except as it refers to what an individual does to solve his own problems and what private and public agencies will do for him in easing that burden." Social prejudice is identified as one of the major problems confronting the handicapped.

The second section contains a discussion of the services that form the basic structure of vocational rehabilitation, namely, physical restoration, vocational guidance, vocational training and selective placement. Vocational rehabilitation would be considerably enhanced if a majority of the medical profession were equally conversant with these fields.

In part three Dr. Kessler describes rehabilitation in practice. Although the author's role is that of an active orthopedic surgeon, his insight regarding the whole man is constant and he has the capacity to convey this perspective to the reader. This section includes discussion of the mentally and emotionally disabled, the orthopedic patient, the blind and the deaf and the medical and surgical invalids.

The final section includes a cursory review of the legislative and administrative organization of a few of the existing programs for the handicapped. The final chapter contains the author's remedies for the problems pointed up so clearly throughout the book. Inadequate rehabilitation is due primarily to "the lack of public and professional knowledge of their possibilities (handicapped) and because of the ignorance of facilities that are already available to them." His major proposal is a uniform, compulsory, lifetime health record in the hands of state departments of health which would require annual reports from the individual or his physician and would urge him "to have his defects corrected by his private physician or by public facilities." Disability pensions are advocated for those who cannot be rehabilitated.

Donald H. Dabelstein

Office of Vocational Rehabilitation
Washington, D. C.

Yoder, Dale. *Personnel management and industrial relations.* (3d ed.) New York: Prentice-Hall, Inc., 1948. Pp. xi+894. $5.00.

For readers familiar with the two previous editions it is sufficient to say that this latest edition maintains the same high quality and thoroughness but is larger, longer, and brought up to date. Addition of materials and developments from the war and post-war period has expanded the discussions on nearly all topics and particularly on selection, wage problems, stabilization of employment, personnel records, and the legal aspects of collective bargaining. When an established authority in the field sees fit to expand his treatment of certain areas it can be used as a rough indication of how the field itself is developing and what problems are receiving major attention.

The major characteristics of this edition are the same as those of the previous editions, viz., recognition of the importance of manpower in our industrial system, growth of personnel management as a profession, treatment of the historical development as well as the present status of the personnel function being discussed, consideration of personnel management problems from many aspects (economic, psychological, sociological, legal), inclusion of a chapter on statistics and reference to appropriate statistical procedures for each topic, emphasis upon research viewpoint and methods, thought provoking exercises and review questions at the end of each chapter, and a thorough, wide coverage of the literature in the field. The literature coverage point should be emphasized since, in addition to footnote references on nearly every page and to collateral readings at the end of each chapter, there is a list of 31 research agencies, 56 journals, and 6 reporting services. The references are quite up to date, nearly all from 1940 on and extending into 1948. The volume is worth its cost just as a bibliographic survey.

A psychologist is in a rather peculiar spot when he reviews a book on personnel management. On the one hand, he likes to see that his psychological viewpoint and findings are permeating the field of management. On the other hand, he doesn't want his findings so thoroughly treated as to eliminate the need for his courses on personnel and industrial psychology. This tendency puts the author of a book on personnel management in a related dilemma. If he doesn't give the psychologist his due, he is criticized; if he includes too much, the psychologist may say, "Stop, you're in my bailiwick."

Yoder has handled this ticklish situation rather well. Of all texts on personnel management with which this reviewer is familiar, Yoder's most clearly reveals the impact of psychological findings upon management principles and procedures, particularly in the areas of selection, training, morale and incentives. The importance of individual differences, interpersonal relationships, and social psychology is stressed in his discussions. Many of his supporting references are from psychological journals. There is still need, however, for a complementary study of the psychological techniques per se and for a thorough treatment of the psychological studies merely referred to in the text. Yoder's text thus will serve both to arouse management's interest in industrial psychology and to help industrial psychologists understand how their contributions fit into the practical situation.

A few minor criticisms could be made, such as the rather poor selection of special ability tests listed as representative of the field (225 n), giving the title of Shartle's book as "Job Analysis" instead of "Occupational Information" (121 n), and stating that the G. I. Bill provides a maximum

of three instead of four years of training (253 n). The only major weakness apparent to this reviewer was the rather casual treatment of supervision.

Albert S. Thompson

Vanderbilt University

Pigors, Paul, and Myers, Charles A. *Personnel administration: a point of view and a method.* New York: McGraw-Hill Book Co., Inc., 1947. Pp. ix+553. $4.50.

This exposition of personnel administration is well organized. Section A (3 chapters) presents the broad function of the personnel administrator, his place in management and the "personnel point of view," based on a recognition of the worker's need for both personal development and social relationships. Section B (3 chapters) presents a method of understanding and solving personnel problems, involving systematic consideration of four elements in the situation: (1) technical features, (2) the human element, (3) principles and policies, (4) the time factor. The use of this "method of situational thinking" by the personnel administrator as a staff officer is described in some detail and the integration of both "person-centered" and "policy-centered" approaches is emphasized. A separate chapter describes the interview as a basic tool in investigation.

Since the personnel point of view stresses individual worker and work-team adjustment and efficiency, Section C (3 chapters) discusses the personnel administrator's function in diagnosing organizational stability through studying employee morale. Indices discussed are production, absenteeism, accidents, turnover, and complaints and grievances. The remaining sections in Part I apply the personnel point of view and method of approach to the standard problems in personnel administration. Twelve chapters deal successively with selection, training, employee rating, transfer and promotion, discipline, wages, hours, employee services, etc. Chapter 22 summarizes the personnel point of view.

The last third of the book (Part II) presents Case Illustrations supplementing the chapter discussions in Part I. Nineteen cases, ranging from 3 to 16 pages each, are given in considerable detail, including background, interview or descriptive data, and interpolated discussion questions and interpretation. Appendices include brief descriptions of the Western Electric Research Program and the Job Relations Training Program of the TWI and a summary of an Employee-Service Program. A Selected References section listing nearly 600 references grouped according to the chapters in Part I, an Index of Names referred to in Part I (but not in the Selected References), and a fairly detailed Subject Index conclude the volume.

This book represents a major contribution to the field and profession of personnel administration. The authors have been able to formulate interestingly and clearly the basic philosophy of personnel work and to show its significance in modern industrial society. It should result in old-line management seeing its problems in a new light and, if studied by beginners, will help create a new generation of personnel administrators alive to their responsibilities. The presentation is particularly strong in its exposition of the staff function of personnel administration, in its guide to the investigation of personnel problems, in its recognition of the inter-relationships between technical and human problems and between person-centered and policy-centered considerations, and in the need for constant appreciation of employee attitudes as a factor in the situation. The crucial position of the supervisor in labor-management relations is stressed and the effect of unionization on personnel practices is evident in the discussions.

The difficulty in evaluating this volume is that it differs from most texts on personnel administration. The sub-title "A Point of View and a Method" describes it nicely for that is just what it does, i.e., it proposes and expounds a frame of reference and a method of approach to the understanding and solution of personnel problems in industry. But, by its very nature, it stops there. Although it is an excellent How to Go About It manual, it is rather weak on What Has Been Done or on How To Do It. There is little attempt to survey the "facts," the "procedures," and the "program" in the standard topics of fatigue, rest pauses, job analysis, job evaluation, labor force characteristics, measurement of employee attitudes, labor laws, employee counseling, personnel record keeping, etc. The Selected References are probably intended to tell the student where to go for this type of information but, if so, the survey is weak in spots, particularly with respect to the contributions of industrial psychologists. The references to psychological literature are mostly textbooks or articles appearing in the AMA publications; only a very few are primary references in psychological journals.

The greatest weakness is an apparent disregard for the method of research in personnel administration. The method of "situational thinking," described so well in Section B and illustrated so consistently in the remaining sections and case examples, is an excellent guide for the handling of the specific case but does not make systematic provision for an organized program of basic research. The research approach, exemplified so well in Yoder's *Personnel Management and Industrial Relations*, is equally important, and, in fact, is necessary to provide the background data upon which the method of situational thinking depends for its validity.

In brief, Pigors and Myers have presented an excellent statement of the personnel point of view and a useful guide for applying its fundamental principles to everyday problems in personnel work. To obtain a well-rounded background for personnel administrators, the student will also need a thorough grounding in research procedures and an extensive factual survey of present knowledge in the field, particularly as revealed in psychological research.

Albert S. Thompson

Vanderbilt University

Doob, L. W., *Public opinion and propaganda*. New York: Henry Holt and Co., 1948, pp. vii-600, $3.75.

To prevent possible disappointment any prospective reader should understand Doob's objective. It was *not* his purpose to review and evaluate the relatively quantitative studies that have been conducted. The principal purpose appears to be an attempt to explain public opinion and propaganda in terms of selected principles of human behavior. Obviously this is quite an undertaking.

In line with this objective, the first group of chapters presents a background and explains such concepts as consistency, rationalization, displacement, compensation, projection, identification, conformity, and simplification.

This is followed by a short outline of "principles of public opinion." The qualifications which must be attached to the set of principles are stated honestly: the concepts which form the basis of the principles are merely characteristics and as such are descriptive only; in view of the uncertain scientific status of existing principles of behavior from which this set of principles has been drawn, it is premature to propose any principles of public opinion and propaganda; the proposed principles need to be extended and refined; and at this stage, "all that principles can accomplish is to call attention to the complexity of the problem and to caution as forcefully as possible against premature generalizations and glibness" (p. 89).

Analyzing results obtained from studies of public opinion was not a principal objective. In fact the "exotic or mundane results obtained from measuring public opinion" are used only incidentally "to indicate the difficulties and the techniques of measurement" (p. iii).

Consequently, the second group of chapters places emphasis on methods of conducting public opinion studies rather than on an analysis of the results obtained. This naturally leads to the consideration of such problems as the nature of the sample; the method of specific assignment (area-type probability sampling) vs. quota sampling; size of sample;

interviewing problems; the technique of questioning; reliability; evaluation of public opinion polls; and such intensive measures of public opinion as panel studies, open-ended interviews, attitude scales, and prolonged (intensive) interviewing.

These chapters evidently represent an attempt to explain the techniques of sampling and measuring public opinion in such simple terms that any one can understand them. The reader who feels the urge to point out what might appear to be inadequate treatment of these subjects should remind himself of the difficulty of reducing the explanation to the simplest possible terms. For example, some readers might object to such statements as "public opinion polls usually draw their samples not completely at random but at random from within specified strata of of the population which have been determined on the basis of attributes related to the particular poll in question" (p. 119). Any one familiar with the way that the most popular polls actually have been conducted might very well question the statement that the selection within strata is really random. However, the point is that the contribution provided by reducing the explanation to a very elementary level probably more than justifies what some readers might regard as inadequate treatment.

There is one point, however, for which the reviewer cannot find a legitimate excuse. In effect, Doob accuses both Gallup and the Psychological Corporation of wording questions to get results which will please their clients (p. 157). Such a statement is so farfetched that it suggests a lack of close practical touch with the ways in which such organizations really operate.

This is merely one example which contributes to the impression that in attempting to cover the numerous specialized fields which make up the general field of public opinion and propaganda, Doob has been forced to rely on reading widely scattered sources rather than depending upon practical experience in each of the fields. His treatment of the field of advertising provides a good example. Obviously his practical experience in this field has been very limited, and his contacts with what has been done in advertising research evidently have not been very close. Yet he did not hesitate to make such statements as: "They (the radio industry) finance polls which purport to show by means of somewhat biased questions that people really like to listen to advertising" (p. 491) in reference to the Field-Lazarsfeld study. The "somewhat biased questions" accusation is neither explained nor supported by any evidence.

The book covers a wide variety of topics in addition to the ones already mentioned, including: the importance of public opinion; the nature of propaganda; such concepts as stimulus intensity, perceptual repetition, perceptual variation, stimulus simplification, reinforcement,

drive reduction, and primacy; the media used for propaganda purposes; and a final summary on the value of analysis, including an outline to serve as a guide in collecting the information needed for a relatively adequate analysis.

What the reader can and cannot learn from the discussion of these topics has been suggested previously. In general, the less the reader knows, the more he will get out of this book. The beginning student will get a relatively quick survey of a wide variety of topics, and the reader who is highly specialized in one field will get at least a surface understanding of the other fields. However, any one with a fairly good grasp of the whole field will find little of interest beyond a few of Doob's personal opinions, and he is likely to feel that the material is fairly thin.

None of these statements is intended as a criticism of the way Doob has approached the difficult problem of covering the whole field of public opinion and propaganda in a single book. To attempt to cover everything from the mechanics of polling to philosophical considerations, with a set of principles included, is a very difficult task. The field is made up mainly of specialists, with each group working in its own specific field and using methods on various levels of accuracy. Coordination is needed. Doob has made a pioneering effort to draw together the scattered threads. For this reason, his book may interest many readers.

<div style="text-align: right">Alfred C. Welch</div>

Knox Reeves Advertising, Inc.
Minneapolis, Minn.

Rudolph, Harold J., *Attention and interest factors in advertising.* New York: Funk and Wagnalls, 1947. Pp 119. $7.50.

For many years Daniel Starch and staff have compiled magazine readership ratings using the recognition method.. The question frequently has been asked, just what do the Starch studies prove? Mr. Rudolph attempts to answer this question, at least in part, and in so doing has presented research findings which throw considerable light on the relative value of a number of present-day advertising techniques. The author states, "the objective of this book was to set forth the elements which contribute to the attention and interest of magazine advertisements and to determine, as far as possible, the extent of each separate influence."

Consumers' reactions to 2,500 different half and full page advertisements appearing in *The Saturday Evening Post* between the years 1935 through 1939 make up the original data for the studies reported. In analyzing these data Mr. Rudolph exhibits an unusual ability to isolate and control a surprising number of factors in advertising. If the book did nothing more than show how such extremely complex data can be

brought under scientific control it would have served a worthwhile purpose; but it does more than that. It produces answers to more than thirty advertising problems, such as relative value of half and full page ads, the best "spot" for a headline, sex preferences for various types of illustrations, and the like.

In certain places one wishes he had isolated a few more factors. For example, he shows the effect of "feeling tone" on attention value in conventional type advertisements but does not show its influence on readership, where one might expect its greatest contribution. In other places even Mr. Rudolph's genius for isolation of individual factors is inadequate and a considerable number of influences operate in an unknown manner. An example of this is his analysis of the problem of "static" vs. "action" pictures. All "static pictures" are lumped together to compare with a similar lumping of all "action pictures" which leaves interest, artistic value, pictorial techniques, and a host of other factors unanalyzed. One must assume that these unanalyzed factors are equally distributed in both types of pictures, which is really a large assumption.

The statistically trained research worker reading this book will find the lack of "N's" and measures of significance of differences a serious shortcoming. The following apology is given by the author: "unfortunately, most of the records pertaining to this investigation were destroyed when the company (J. Stirling Getchel, Inc.) went out of business. For this reason it is not possible to show the number of advertisements involved in each separate comparison." It is unfortunate such valuable data were destroyed.

While the author emphasizes the specificity of the problems dealt with, it may be well to stress further the fact that the techniques investigated are concerned primarily with the mechanical aspects of advertising. If one believes mechanical perfection will make a successful advertisement, then this book is an extremely important contribution to advertising. If, however, one follows Kenneth Goode, H. C. Link, and others who hold that mechanical factors, while important, are decidedly secondary to the advertiser's ability to tap deep undercurrents of human motives, then the importance of this book is whittled down considerably.

Howard P. Longstaff

University of Minnesota

Selekman, Benjamin M. *Labor relations and human relations.* Cambridge, Massachusetts: McGraw-Hill Book Company, 1947, pp. xi + 225.

Another appropriate title for Selekman's book would be *A Psychology of Labor Relations* and it is indeed a reflection upon psychologists as a

group that one of them has not come forward to write a volume on this subject. Granted that a sizable body of empirical facts has not been developed in this area, it is heartening to see someone strike out and prepare an exploration of the human relationships involved in negotiating and living under a union agreement. While other authors have discussed the subject, this organized treatment is long overdue.

Selekman paints the current picture of strife in the world of union relations and raises the questions of "why" and "what can be done about it?" "How can we achieve in daily shop behavior the cooperation necessary for realizing both full production and maximum human satisfaction?" His answer traces the emotional reactions of both union and management men from the time a union enters the industrial scene as an organizing unit, through the negotiation of the first agreement, to the problems of administering and modifying the agreement. His plea is for greater common understanding of the person across the bargaining table as a human being with foibles and feelings, motivations and frustrations. He insists that "the capacity for conflict *and* cooperation lies deep in the human endowment." Conflict is today's pattern because modern industrial organization and local shop practices tend to hide the realities of interdependence which usually build spontaneous cooperation. "The discovery of methods for imparting to each man at work the feeling that he is an indispensable part of the whole working group thus figures as a major problem for research and experiment." Selekman faces squarely the very real bases for disagreement and suggests means of meeting the problems thus arising. It is interesting to relate his proposed practices with those found by the National Planning Association in its series, *Causes of Industrial Peace.*

The psychologist will find many challenging questions; the last chapter, "Conflict and Cooperation," presents several hypotheses which could serve as effective foundations for research. While he may not agree entirely with all of the interpretations, he will be stimulated to do some thinking in a much-neglected (by him) field. The book could be useful to the industrial psychologist for distribution to his friends in labor relations and in unions; they may find it a trifle hard to read but it will be well worth the effort. It is a book that the non-industrial psychologist will find valuable in understanding the potential role of the psychologist in union relations.

Brent Baxter

Personnel Department
The Chesapeake and Ohio Railway Company
Cleveland, Ohio

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to
Donald G. Paterson, Editor, Department of Psychology, University
of Minnesota, Minneapolis 14, Minnesota

You and your mental abilities. Lorraine Bouthilet and Katharine Mann Byrne. Chicago: Science Research Associates, 1948. Pp. 48. $.75 single copy. $.60 for fifteen or more. $.40 for one hundred or more.

The psychology of social classes. Richard Centers. Princeton: Princeton University Press, 1949. Pp. 256. $3.50.

The ethics of ambiguity. Simone De Beauvoir. New York: Philosophical Library, Inc., 1948. Pp. 163. $3.00.

The people know best. Morris Ernst and David Loth. Washington, D. C.: Public Affairs Press, 1949. Pp. 169. $2.50.

The psychology of invention in the mathematical field. Revised edition. Jacques Hadamard. Princeton: Princeton University Press, 1949. Pp. 145. $2.50.

Elmtown's youth. A. B. Hollingshead. New York: John Wiley and Sons, Inc., 1949. Pp. 420. $3.50.

Conference guide to basic management training. Arthur S. Hotchkiss. Deep River, Conn.: National Foremen's Institute, Inc., 1949. Pp. 206. $5.50.

Understanding yourself. William C. Menninger. Chicago: Science Research Associates, 1948. Pp. 52. $.75 single copy. $.60 for fifteen or more. $.40 for one hundred or more.

Historical introduction to modern psychology. Revised edition. Gardner Murphy. New York: Harcourt, Brace and Co., Inc., 1949. Pp. 466. Textbook $4.50, Trade $6.00.

Machine computation of elementary statistics. Katharine Pease. New York: Chartwell House, Inc., 1949. Pp. 238. $2.75.

The pollsters. Public opinion, politics and democratic leadership. Lindsay Rogers. New York: Alfred A. Knopf Company, 1949. Pp. 239. $2.75.

Music and medicine. Dorothy M. Schullian and Max Schoen, Editors. New York: Henry Schuman, Inc., 1948. Pp. 499. $6.50.

Intellectual abilities in the adolescent period. David Segel. Bulletin 1948, No. 6, Federal Security Agency. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1948. Pp. 41. $.15.

How personalities grow. Helen Shacter. Bloomington: McKnight and McKnight, 1949. Pp. 256. $3.00.

Human relationships in public health. Geddes Smith. New York: The Commonwealth Fund, 1949. Pp. 18. $.15.

The American soldier: Vol. 1. Adjustment during army life: Vol. 2. Combat and its aftermath. S. A. Stouffer et al. Princeton: Princeton University Press, 1949. Pp. 600 each. Vol. 1 and 2, $13.50. Separate, $7.50.

Appraisal of vocational fitness by means of psychological tests. Donald E. Super. New York: Harper and Brothers, 1949. Pp. 780. $6.00.

Dynamic psychology. Percival M. Symonds. New York: Appleton-Century-Crofts, Inc., 1949. Pp. 413. $3.75.

Personnel selection. Test and measurement techniques. Robert L. Thorndike. New York: John Wiley and Sons, Inc., 1949. Pp. 366. $4.00.

Perspectives in medicine. New York Academy of Medicine. New York: Columbia University Press, 1949. Pp. 163. $2.50.

Research frontiers in human relations. Vol. 92, No. 5 of Proceedings of the American Philosophical Society. Philadelphia: American Philosophical Society, 1948. Pp. 86. $1.00.

How to prepare an employee's handbook. Deep River, Conn.: National Foremen's Institute, Inc., 1949. Pp. over 300. $12.50.

The new cure for white collar unrest. New York: Prentice-Hall, Inc., 1948. Pp. 47. $1.00.

A Vocational Interest Test at the Skilled Trades Level *

Kenneth E. Clark

University of Minnesota

The counseling of college students who plan to enter one of the professional fields has been greatly aided by the development of the Strong *Vocational Interest Blank* and the Kuder *Preference Record*. Widespread use of these devices not only with college students but with high school students, job applicants, the unemployed, and other groups has demonstrated the usefulness of a measure of an individual's interests in comparison with those of successful workers in a given occupation. One of the serious limitations of these instruments, however, is the inadequate coverage of occupations at the skilled and semi-skilled levels. Thus, the Strong *Vocational Interest Blank* can be scored only for carpenter, printer, and policeman at these occupational levels.[1] As a result, the vocational counselor is much better prepared to counsel the small minority of potential professional, semi-professional and technical workers than to counsel the large majority of persons planning to enter skilled, semi-skilled, and unskilled occupations—at least as far as the measurement of interest patterns is concerned.

During World War II, the armed services placed great emphasis on the measurement of aptitudes; little was placed on the measurement of interests. It frequently happened that, when highly capable men were sent to technical schools for training, school officials would often complain that the students were not "interested." That it would have

[1] The trade unions used in this report are members of the St. Paul Trades and Labor Assembly, and include Bakery and Confectionery Workers, No. 21; Electrical Workers, No. B-110; Milk Driver Employees, No. 546; Painters, No. 61; Plasterers and Cement Finishers, No. 20; Plumbers, No. 34; Sheet Metal Workers, No. 76; Typographical Union, No. 30; and Steam Fitters-Pipe Fitters, No. 455.

been desirable to pay more attention to measured interests of individuals was generally recognized. To actually do so, in practice, was rather difficult. For one thing, military terminology is strange to the newly inducted recruit. To ask for statements of job preferences in terms of job titles is therefore likely to be futile. To ask for a statement of preferences in terms of definite types of activities is also likely to obtain information of doubtful value either from a civilian or a military respondent. Even were such an approach considered desirable, it is likely that the high level of affect among recruits would lead them to state preferences in terms of assignments which either keep them closer to home, keep them in the continental United States longer, or either reduce or increase their likelihood of being assigned to combat duty. The use of a questionnaire which could be scored to indicate the interests of an individual in terms of the known interest patterns of members of military occupational groups was not possible because such an instrument did not exist. It is the purpose of the present investigation to explore the possibilities of developing such an interest measure, usable for potential workers both in the occupations of enlisted men in the armed services, and in the corresponding civilian occupations.

The Questionnaire

To provide the information on preferences needed for the analysis of interest patterns, a 570-item questionnaire, the *Minnesota Vocational Interest Inventory*, was prepared. Items in the questionnaire were grouped in three's, making up a total of 190 triads. The individual respondent is asked to select from each triad of items the one activity he would like most, and the one he would like least, leaving the third item blank. The approach used is thus a forced-choice, with the respondent who follows directions being required to make a total of 380 choices, half of them "like" and half of them "dislike."

The items used in the inventory were selected from a variety of sources. First, a large number of items were written which described jobs or tasks making up part of a job.[2] The *Dictionary of Occupational Titles*,[3] the *Manual of Navy Job Classifications* (Nav Pers 15105)[4] and similar materials were scrutinized for suggestions for items. The final list of items used contains such activities as the following, grouped in three's as shown, with the directions for marking responses as indicated:

Directions

On the following pages you will find many activities listed. They are arranged in blocks of three. You must make a choice in each block of the one thing you like to do most, and the one thing you like to do least.

Mark the thing you like to do most with a plus-sign (+).
Mark the thing you like to do least with a minus-sign (−).
That leaves one of three items blank.

> Example: ( ) a. Write letters.
>
> (+) b. Fix a leaky faucet.
>
> (−) c. *Interview someone* for a newspaper story.

Now turn the page and begin. Be sure to fill out *all* pages.

Items are grouped in three's in a haphazard fashion. Thus, no *a priori* plan of scoring played any role in determining how items were combined to make triads. As a result, the blocks of three look like the following examples:

a. Be a grocer.
b. Be a printer.
c. Be a shop foreman.

a. Varnish a floor.
b. Learn to use a slide rule.
c. Repair a broken connection on an electric iron.

a. Tune a piano.
b. Cook a meal.
c. Change a tire on an automobile.

a. Putter around in a garden.
b. Take part in an amateur contest.
c. Cook spaghetti.

It should be noted, however, that although items are not grouped in three's in any pre-arranged manner with regard to possible responses of individuals in different occupations, nonetheless some attempt was made to keep the nature of the items within the same order of complexity. Thus, learning calculus is not combined with an item on cooking a meal, since these two items are not compatible in terms of the ordinary life situations which would present a choice between two such alternatives. Even this kind of control did not always operate, so that one discovers such an odd combination as the following in the finished questionnaire: a. Address envelopes; b. Try to find an error in a financial account; and c. Help put out the fire in a burning building.

Several difficulties are encountered in an approach of this sort. The obvious one, and the one which produced most frequent comments from the skilled tradesmen who cooperated in this investigation, is that the combination of forced-choice and haphazard arrangement of items in groups produced many triads in which a decision is difficult. Thus, many highly masculine members of the electrician's group had difficulty finding an item to mark with a plus ("like") in the following group: a. Put a closet in order; b. File cards in alphabetical order; and c. Make a pie. In spite of this difficulty, the writer believes this method is still to be preferred to the more obvious type of choice which is made when a questionnaire's content is "stacked."

Decisions regarding types of items, and mode of response, were made largely on the basis of rather meager experimental evidence, and on the basis of subjective appraisal of the types of items which would work best in the situations where it is expected the inventory will be used.

The Criterion Groups

The plan of the present inquiry required that the questionnaire be administered to successful employed workers in the skilled trades occupa-

tions. It was not considered desirable to prepare keys solely in terms of rubrics identified by any other means.

The first contacts with employed groups were made through various business and industrial organizations in Minneapolis and St. Paul, Minnesota.[5] Willingness to cooperate in the project was expressed by personnel managers of many of these concerns, with the reservation that the matter should be cleared with union representatives before any action was taken. Furthermore, some reluctance to use company time for the collection of data was expressed, along with assurances that this, indeed, was a worthy project.

Union representatives were, accordingly, contacted, and the possibilities of the program described to them. Union leaders were, on the whole, willing and anxious to cooperate in any program which might eventually operate to increase their own effectiveness in selecting apprentices for training in their own trades. With only one exception, union groups who were contacted agreed to aid in the assay of interest patterns of their own memberships.

How to obtain the responses of the membership still remained a major problem. The first attempt was attendance at union meetings. The program of research was presented to the membership with a request for their cooperation in responding to the questionnaire during the meeting itself. Inasmuch as completing the questionnaire required from 45 minutes to over an hour, this effort proved to be futile. Somehow or other union meetings were not considered by the membership a suitable time for this sort of work, and as a result, many questionnaires were begun, but few were completed.

A second effort was made at the place of employment. A representative of the project, accompanied by a union official, would make the rounds of places of business, would describe the research program to the worker, who would then fill out the questionnaire while our representative and the union representative waited. This method was most effective, although unpopular with the employer, and excessively time-consuming.

A third attempt was by use of the mails. Through the trust and cooperation of the union leaders, it was possible to use their mailing lists, and to send questionnaires to the membership, with a covering letter signed by the business agent, or another official of the union. These letters asked for the cooperation of the membership, gave a brief word about the objectives of the program, and gave the endorsement of the local union officials to the project. A stamped envelope was enclosed, addressed to the union office. Returns were anonymous, with questionnaires coded in such a way as to permit follow-ups only to those who had not yet returned their questionnaires.

Of the various methods tried for obtaining adequate coverage of workers in a given occupational group, the mail questionnaire method proved to be most effective. Thus, while 3500 questionnaires were distributed in union meetings, with the membership voting the program "hearty support," only 129 usable returns were obtained. Mail questionnaires to 320 electricians yielded, with one follow-up letter, a return of 201 questionnaires—a 63 per cent return, of which most were usable. For the A. F. of L. trade unions used so far, questionnaires have been mailed to the entire male membership. Figures on percentages returned and usable are shown in Table 1.

The sampling of occupational groups for the purpose of describing interest patterns requires that the segment used be representative of the entire adult group employed in the field. To what extent is this true of our groups? Several factors operate to bias our samples:

1. Geographically, the samples are highly restricted. Only if the St. Paul skilled tradesmen are strictly representative of the skilled tradesmen in the same occupations all over the country will this bias be eliminated. It is intended that this geographic bias be reduced in later samples, not only by securing returns from Minneapolis workers, but by obtaining samples in other localities.[6]

2. Only skilled tradesmen who are members of St. Paul locals of the American Federation of Labor are included in the samples. This, obviously, is a serious source of bias, and one which will need to be corrected. This study is restricted to the A. F. of L. unions partly as a matter of expediency, and in part because these unions are trade unions, not industrial unions. Working with limited funds, this study could not even exploit all of the data-collecting opportunities provided by the St. Paul Trades and Labor Assembly of the A. F. of L., and so there was little point in diversifying the required contacts.

3. All members of a particular union local did not return their questionnaires, and so were never included in a sample. While a 100 per cent sample would have been ideal, it becomes much too expensive to even try. As noted in Table 1, the coverage of a particular union is in each instance better than fifty per cent, but never more than 75 per cent.[7] It is likely that those workers who responded to our mail appeals represent a different type of person from those who did not respond. How serious a bias this is cannot easily be assayed.

It may be that these biases which affect all of the groups do not influence the difference between groups, since the preparation of occupational keys requires the determination of the interest patterns which *differentiate* one occupation from another. Thus, geographic location may produce a larger number of workers who say that they like to fish, but would not affect the size of the differences *between* groups in this expression of interest.

The present report concerns itself with an analysis of the responses to the *Minnesota Vocational Interest Inventory* of workers in the eight civilian occupational groups for whom samples of some size are available. These groups, and their sizes, are reported in the last column of Table 1.

Development of a Tradesmen-In-General Group

In order to score the responses of men in a given occupational group to show the items which are answered in the same way by these men, it is necessary to have some basis of comparison with persons not in that occupational group. Thus, it is not enough to know that 75 per cent of electricians respond to an item in the same way, for it may be that 75 per cent of all men would respond in that way. To obtain a group of adults who would represent a cross-section of all adult men in the skilled trades has not been possible within the scope of this project. To obtain an *estimate* of the responses which such a group would make, the following procedure was employed:

[6] During World War I, and subsequent work of the Occupational Research Program Staff of the U. S. E. S., trade tests were standardized in three geographically separated localities to overcome possible local peculiarities in trade practices.

[7] Information given in Strong's *Vocational Interests of Men and Women*, does not indicate what percentage returns he attained. However, it is probable that he seldom achieved a 75 per cent return, or even a 50 per cent return.

a. The percentage response of members of each of the eight occupational groups in the civilian trades being analysed (listed in Table 1) to every item of the inventory was computed.

b. The percentage responses for each of the eight groups were added together and divided by eight, giving the *average* percentage response to each item.

c. This average percentage response was used as the best estimate of the percentage response which would be made by members of an actual tradesman-in-general group.

It should be noted that this procedure gives the identical values which would have been obtained if a representative and equal sample of the respondents in each of the eight civilian occupations had been used to make a single total group. The procedure operates to eliminate the over-weighting of occupational groups with a large number of members, and the under-weighting of occupational groups of small size.

This estimate is obviously not an entirely satisfactory solution to the problem of getting a tradesmen-in-general group which is truly representative. However, it seems likely that for the purposes for which the interest inventory is being developed, the procedure gives an adequate base for *preliminary* comparisons between groups.

Table 1

Numbers of Questionnaires Mailed and Returned for the Eight A. F. of L. Unions Sampled and Numbers Used in Developing Scoring Keys

A. F. of L. Union	Number Sent	Number Returned	Per Cent Returned	Number Usable	Per Cent Usable	Number Used for Keys*
Electricians	320	201	63	166	83	185
Milk Wagon Drivers	608	326	54	218	67	127
Painters	712	390	55	267	68	252
Plasterers	167	111	66	74	67	51
Bakers	473	305	64	144	47	64
Sheet Metal Workers	298	220	74	164	75	99
Printers	530	331	62	278	84	300
Plumbers	576	347	60	199	57	65
					Total	1143

* N in this column is the number of persons in the different unions who identified themselves as belonging to a particular occupational group. Thus a few electricians were found in unions other than the electricians's union itself; within the milk wagon drivers' union were workers who were not actually milk wagon drivers.

The Preparation of Scoring Keys

The purpose of a scoring key for a particular occupation is to make possible the comparison of responses of an individual to the responses of members of a given occupational group, to determine whether or not the individual's responses are like or unlike those of a given group. It is necessary, therefore,

to compare the responses of the group to the responses made by the tradesmen-in-general group. Consider the example below:

Item 18.	Percentage Responses of "Like" Made by:	
	Tradesmen-in-general	Electricians
a. Be an electrical engineer	61%	90%
b. Be an aeronautical engineer	20%	5%
c. Be a surgeon	19%	5%

It is apparent that electricians have, as a group, a more general preference for being an electrical engineer than do tradesmen-in-general, although even the latter group selects this response more frequently than either of the other two responses. We may score a response of "like" to the response *a.* as a response counting towards a high score on the electricians' key, since a significantly larger proportion of electricians pick this response than do the tradesmen-in-general. A response of "like" to either of the other two items, however, may be scored as a response detracting from a high electricians' score, since a smaller proportion of electricians pick this response than do tradesmen-in-general. Thus, we might make up our electrician's key as follows:

a. Be an electrical engineer   A response of + counts 1 point
b. Be an aeronautical engineer   A response of + counts −1 point
c. Be a surgeon   A response of + counts −1 point

However, the respondent has also selected one of these three items as the one which he likes least, or dislikes most, and has marked that item with a − mark. Therefore these percentage responses must also be scrutinized.

Item 18.	Percentage Responses of "Dislike" Made by	
	Tradesmen-in-general	Electricians
a. Be an electrical engineer	12%	1%
b. Be an aeronautical engineer	21%	21%
c. Be a surgeon	67%	78%

So in similar fashion, these responses may be scored as follows:

a. Be an electrical engineer   A response of − counts −1 point
b. Be an aeronautical engineer   A response of − counts 0 points
c. Be a surgeon   A response of − counts 1 point

Thus, to generalize—a response made more frequently by the members of a particular group than by the tradesmen-in-general group is scored as a plus in the key for that particular group. A response made less frequently by the members of a particular group than by the tradesmen-in-general group is scored as a minus in the key for that group.

How shall the blank item be scored? It is apparent that such a response has meaning, just as the indifferent response on the Strong *Vocational Interest Blank* has meaning. The present analysis, however, ignored the blank item in scoring for two basic reasons: 1. percentage responses of like and dislike reflect the effects of leaving items blank, and therefore use the blank response in scoring, and 2. the scoring of the absence of a pencil mark is a complicated task in hand-scoring, and an almost impossible task in machine scoring.

How great a difference should be required? This question cannot be answered with finality. In this report, an 11 per cent difference was used as the minimum value required for using an item response in a given key. This value represents a compromise between the 6 per cent value used by Strong

and the 20 per cent and higher values used with success by Hathaway and McKinley[8] in the development of the *Minnesota Multiphasic Personality Inventory*.

Should greater percentage differences contribute more to total score than smaller ones? The Strong *Vocational Interest Blank* increases the differentiation between groups by assigning greater weights to responses which differ markedly in the criterion group from the responses made by men-in-general, and assigning smaller weights to responses where the difference is smaller. Preliminary data, in this study, however, indicate a possibility that use of multiple weights is not required to maximize the differentiation of occupational groups.

Comparisons on the Eight Occupational Keys

All members of the eight occupational groups were scored on the key for their own occupation. The distributions of these scores are listed in Table 2. Also presented in Table 2 are distributions of scores on each of these keys of persons not employed in the occupation. A comparison of each pair of distributions gives an indication of the degree to which the occupational scoring keys actually work in separating workers in a given field from persons in other skilled trades jobs.

The distributions of scores of workers outside the occupation were obtained by scoring a sample of 25 inventories of workers from each of the other seven occupations. The selection of inventories was made on a random basis. A comparison of the distribution of scores of these samples of 25 with the distributions of the total group on its own occupational key showed only slight differences. The scores of each of the combinations of seven groups (N = 175) not belonging to a given occupation are distributed in Table 2 as the scores of tradesmen-in-general.

Marked differences exist between the distributions of scores of non-members of an occupation and those of employed workers in the occupation. Since the primary purpose of this investigation is to determine whether or not it is possible to separate members of skilled trades groups on the basis of their measured interests, measure of overlap between distributions of members and non-members is the appropriate statistic. Table 2 presents the per cent of the tradesmen-in-general exceeding the median of a given skilled group. This value varies for the different keys from 2.0% to 14.3%, with a median value of 6.3%. Thus, about six out of 100 workers not in a give occupation make scores above the median of employed workers on the typical scoring key prepared in this study.

The percentage 6.3 does not compare too favorably with the values of two or three per cent obtained by Strong in his work with his *Vocational Interest Blank* at the professional levels. Does this difference indicate that trades groups are more nearly alike than are professional groups,

[8] S. R. Hathaway and J. C. McKinley. *Minnesota Multiphasic Personality Inventory, Manual.* New York: The Psychological Corporation.

Table 2

Distributions of Scores of Workers in a Trade (Group A) and Tradesmen-in-General (Group B) on Each of Eight Occupational Scoring Keys

Score	Plasterers A	B	Milk Wagon Drivers A	B	Printers A	B	Electricians A	B	Painters A	B	Bakers A	B	Sheet Metal Workers A	B	Plumbers A	B
120–129							4									
110–119							26	1							4	
100–109							46	4							18	
90–99							38	14					13	1	12	9
80–89							26	20					26	12	13	17
70–79							16	18					19	21	7	27
60–69					1		12	16					20	25	3	28
50–59	1				5		7	20			1		10	30	6	21
40–49	8				21		5	24			1		4	27		18
30–39	16	6			21	2	4	14			1	1	3	19	2	21
20–29	16	29	3	1	36	3		18	34		6	1	3	14		10
10–19	7	52	15	1	51	3		12	127	5	6	1	10			10
0– 9	1	49	29	9	39	8		4	73	65	5	3	7			5
−10−−1	2	30	31	14	41	15		7	17	91	10	4	1	5		6
−20−−11			22	19	44	27	1	1	1	14	11	14		4		3
−30−−21			16	30	25	22			1		6	12				
−40−−31			9	35	9	41			1		9	29				
−50−−41			2	40	7	36					8	28				
−60−−51				23		16						30				
−70−−61				3		2						21				
−80−−71												27				
−90−−81												4				
N	51	175	127	175	300	175	185	175	252	175	64	175	99	175	65	175
Mdn.	30	10	−5	−34	6	−32	96	82	13	−2	−12	−48	74	51	91	57
No of Items in Key	131		144		168		226		66		212		168		211	
Per Cent Overlap*	3.4		10.3		6.3		6.3		2.0		7.4		14.3		4.6	

* Per cent of distribution of scores of tradesmen-in-general exceeding median of the distribution of scores of members of a trades group.

or is there another explanation? The writer believes that the use of a constant eleven per cent difference for deciding to use or not use a given differential response in the scoring key does not give maximum separations between criterion and control groups. This factor, plus the fact

that multiple weights were not used in the present keys, but are used by Strong, operates to increase percentage overlap.

Inspection of Table 2 reveals that the eight keys differ not only in their power to separate members of a trade from outsiders, but also in the kinds of distributions which they produce. One is immediately struck by the differences in the median raw scores attained by workers on their own occupational keys. Electricians have a median score of *plus* 96, while bakers have a median score of *minus* 12. It is easy enough to see how high plus scores are attained, but why should a group get a minus score on its own key? The answer is to be found by studying the percentage responses of bakers to the items which differentiate them from tradesmen-in-general. These item responses are generally unpopular items. Tradesmen-in-general select them only 10 to 20 per cent of the time; even bakers select each item, on the average, less than 50 per cent of the time. This is a rather interesting finding, since it indicates that interest patterns may operate to differentiate occupational groups not only by use of items selected by an overwhelming majority of a group, but also by use of items actually rejected by a majority of a group.

The differences in variability of the distributions of scores reflects to a considerable extent the number of responses included in the scoring key. Since there are 570 items in the inventory, and since either response of like or dislike is scored, a total of 1140 responses are scorable. Whereas a large number of items differentiated electricians from non-electricians (226), only a small number of items did the same job for painters and non-painters (66). This difference is undoubtedly due, in part, to the kinds of items included in the inventory. It may also be due, in part, to real differences in the degree to which workers within an occupational group resemble each other. It is possible that painters, as an occupational group, have fewer basic interests in common than do electricians.

A small dispersion of scores is usually associated with lower reliability, which suggests that the keys with less variability are also those with high percentages of overlap between criterion and control groups. A comparison of the percentage overlap with the number of items in the key, as given in Table 2, gives no support to this notion. The present data indicate that a small dispersion of scores is not, in itself, undesirable.

Relationships Between Scoring Keys

Table 3 presents the correlations between scores on each of the eight keys obtained for the sample of 200 tradesmen whose inventories were scored on all keys. These correlations range from high positive to high

negative values, giving the impression that the scores on various keys tend to cluster in rather meaningful patterns. Thus, the interests of milk wagon drivers and bakers seem to have much in common, as do the interests of electricians, sheet metal workers, and plumbers. The interests of painters, on the other hand, have little relation to those of any of the other seven groups.

The small number of occupations involved makes it difficult to determine to what extent these clusterings result from the sampling of occupational groups in the study, and to what extent they result from real similarities of interest of the different workers. It is easy to see how milk wagon drivers and bakers would appear very much alike when compared with workers in the building trades; whether or not this same degree of relationship would hold if the sample of occupations were larger,

Table 3

Intercorrelations of Scores on Eight Occupational Keys for a Sample of 200 Tradesmen*

	Plasterers	Milk Wagon Drivers	Printers	Electricians	Painters	Bakers	Sheet Metal Workers	Plumbers
Plasterers		−.11	−.60	.12	.30	−.21	.38	.26
Milk Wagon Drivers			.45	−.73	.03	.84	−.77	−.78
Printers				−.68	−.06	.58	−.81	−.74
Electricians					−.19	−.83	.85	.85
Painters						.04	−.12	−.13
Bakers							−.85	−.86
Sheet Metal Workers								.89
Mean	11.2	−28.6	−24.1	56.5	−0.5	−41.9	49.3	56.4
Standard Deviation	13.35	19.00	22.85	32.89	7.80	26.17	25.56	23.14

* 25 men from each of the eight occupations were scored on all eight keys. Thus, each key is scored for 25 members and 175 non-members of the given occupation.

and more heterogeneous is not clear. It is fairly certain that the actual separations of workers from outsiders achieved in this study would have been considerably more spectacular if a wider diversity of occupations had been included. The degree to which electricians, plumbers and sheet metal workers cluster, as do bakers and milk wagon drivers, tends to obscure the marked differences between the two clusters and the remaining occupations.

Another method of portraying the clustering of occupations is used in Table 4, in which the median percentile score of each of the eight groups of 25 is given for two keys—the electrician's key and the milk wagon driver's key. Percentile scores are computed on the distributions of scores of

Table 4
Median Percentile Scores* on Each Occupational Key Attained by 25 Electricians and 25 Milk Wagon Drivers

Key	25 Milk Wagon Drivers	25 Electricians
Plasterers	8	5
Milk Wagon Drivers	48	1
Printers	5	3
Electricians	3	51
Painters	7	5
Bakers	26	1
Sheet Metal Workers	7	25
Plumbers	3	22

* Percentile scores are computed using the distribution of scores of each group of employed workers on its own occupational key; that is, the distributions of scores for group A listed in Table 3.

workers in the occupation. The close relationship between the milk wagon driver's and the baker's key is indicated when the median score of milk wagon drivers on the latter key is at the 26th percentile of the baker's distribution of scores. The same sort of clustering occurs between electricians, sheet metal workers, and plumbers.

Summary

The present report has analyzed the interests of members of eight A. F. of L. trade unions. When scoring keys are prepared to differentiate between members of a trade group and a composite group of tradesmen-in-general, it is found that:

1. Workers in a trade can be separated from workers in other trades on the basis of their measured interests with considerable success. About six workers out of a hundred will exceed the median score of tradesmen in an occupation other than their own.

2. The separation is achieved with a rather crude criterion for preparing scoring keys: that the response of the one group differ by eleven percentage points or more from the response of the composite tradesmen-in-general group.

3. Distributions of scores on the different scoring keys vary considerably both in central tendency and in variability, but these values are not closely related to the goodness of the keys, as defined by the degree of separation between workers in and not in the trade.

4. Correlations between scores on the eight keys indicate a clustering of trades with respect to measured interests. Workers in three unions

related to the building trades (electricians, plumbers, and sheet metal workers) tend to have related interests, but to differ markedly both from workers in two service occupations (milk wagon drivers and bakers), and from workers in two other building trades (painters and plasterers).

5. The data analyzed thus far seem to suggest that skilled trades groups may be ordered into families of occupations with rather similar interests, so that it may not be necessary or desirable to differentiate between closely related occupations either in preparing separate scoring keys, or in the guidance of young persons contemplating entry into these fields of work. However, this aspect of this program of research requires considerably more work than has been completed thus far.

A Selection Battery for Bake Shop Managers *

Edwin B. Knauft

Federal Bake Shops, Inc., Davenport, Iowa

A number of investigators (1, 3, 6, 11, 16, 18) have attempted to develop and validate series of items or test batteries which would efficiently predict executive or supervisory job success. Some of these studies were moderately successful, but the majority reported "validity" data which were based only on the original population used in the standardization or item analysis of the tests. It is generally recognized that the abilities or characteristics contributing to success variance in managerial positions are difficult to isolate and measure. The problem is further complicated by the fact that it is often impossible to obtain a relevant and reliable criterion of supervisory job success. In addition, validation is difficult because it is unusual for a large number of supervisors or managers to be engaged in the same or similar job duties.

The objective of the present study is to construct and attempt to validate a series of written tests which will predict subsequent on-the-job behavior of shop managers in a retail—manufacturing bakery chain. Seventy-nine managers of bake shops were available for the initial research, 85 managerial applicants were used in the subsequent development of test norms and 33 new managers were followed up on the job and formed the cross validation study.

The Bake Shop Manager

The managers were employed by a chain which operates 88 retail-manufacturing bake shops in the Midwest, East and South. Each shop is under the supervision of a manager who directs both the manufacture of bakery products from raw materials and the sale of the products to the public.

The principal duties of the manager may be summarized as follows: (1) purchases raw materials; (2) directs and sometimes participates in the manufacture of baked products from these raw materials; (3) determines the variety of products and quantity of these products which shall be produced each day; (4) computes the cost and determines the selling price of each product; (5) hires, discharges and supervises the work of bakers, baker apprentices, sales-

girls and porters; (6) keeps financial records, pays employees and writes checks for raw materials purchased; (7) sends financial, sales and inventory reports to the home office; (8) works under the general supervision of a district manager.

The Criterion Problem

Since each manager has primary responsibility for the successful operation of his unit, it is reasonable to presume that the general financial condition of the unit—in terms of profit and loss—would reflect the abilities of the manager. The actual profits of each unit, however, are not a satisfactory measure of managerial ability because certain expenses not under the manager's control affect the profits. The standard company accounting procedures divide the expenses of each unit or shop into controllable and uncontrollable costs. Variables largely under the control of the manager are grouped together and are known as total controllable costs. The actual dollar volume of these costs is partly a function of the total sales of a unit, and hence a direct comparison of those figures from unit to unit would place the manager of a small unit at considerable disadvantage. For this reason, the ratio of this cost to the total sales of the unit is computed and affords a unit to unit comparison. This measure, designated hereafter as total controllable cost, is one possible criterion measure of managerial ability.

Data on the total controllable costs were obtained for all units from the 1946 Company operating statements. The data were first analyzed by districts and it was found that there were rather large differences between the means of certain districts. An analysis of variance was made of these district data to test the hypothesis that the district means varied from each other only by chance. The resulting F value of 4.92 is significant at better than the 1% level of confidence, indicating that these differences may be due to factors other than chance. It therefore seemed reasonable to use the controllable cost ratios as a measure of individual manager performance only after these data had been corrected for the "district effect." A district correction factor was applied to the 1946 cumulative cost percentages of each unit. The base for this correction factor was the difference between the 1946 controllable cost for the entire company and the corresponding value for the given district.

The relationship between length of time in a managerial position and the above corrected measures was investigated to determine the effect of experience. It was found that nine men who had been managers from three to six months did not have a mean corrected controllable cost percentage which was significantly different from the mean of similar measures for 61 managers who had been on the job for more than one year. It therefore seemed reasonable to include in the original study all managers who had been on the job for three months or more.

The corrected reliability of the corrected controllable cost data, by units, as estimated from the values of odd vs. even months, was .96.

Several members of the management of the Company felt that the raw materials cost should be given more weight in a composite criterion than was reflected by the actual contribution of raw materials to total controllable cost. There was available a raw materials percentage (ratio of dollars spent for raw materials to dollar sales of the unit) which reflects the manager's ability to buy raw materials wisely, to prevent waste during production and, indirectly,

to correctly set the selling prices of his products. The raw materials percentage for each unit for 1946 was used as a second criterion measure. These data were analyzed by districts in the same manner as the total controllable costs and again the analysis of variance yielded an *F* value (3.14) which was significant at better than the 1% level of confidence. A correction factor computed in a manner similar to that described above for controllable costs, was calculated on the basis of the difference between the raw material percentage for each district and the raw material percentage for the entire company. The correction factor for each district was then applied to the raw material percentage of each unit in the given district. The reliability of the raw material measure was estimated by correlating the corrected unit data of odd months of 1946 with even months. The resulting corrected coefficient was .92. These reliability data were based on 63 units in which there was no change in managers during the year.

A Subjective Rating of Performance. The manager's job is so complex that many aspects of managership probably are not directly reflected in the two criterion measures which have been discussed. Some type of merit rating procedure appeared to be the only technique which would measure those managerial qualities not reflected in financial data of the units. A survey of various types of personnel merit rating methods (9) led to the conclusion that the weighted check-list type of rating scale would be appropriate in the present situation. This technique, involving the equal appearing intervals method of Thurstone (17), was first applied in a merit rating situation by Richardson and Kuder (14).

The procedure used to construct a weighted check-list rating scale for bake shop managers has previously been reported in detail (10). This scale, which was comprised of two forms of 24 items each, was used to evaluate the 79 managers in the initial study. The basic rating data were obtained from the evaluations resulting when each district manager applied both forms of the scale to his unit managers. The product moment correlation of scores obtained on the two forms of the scale was .79. The reliability of the combined scale consisting of both forms was estimated by the Spearman-Brown formula to be .88. Additional data were available on 35 of the managers who were also rated by their respective assistant district managers. The reliability coefficient of the scale, based on the ratings of the 35 managers by two superiors was .81.

The rating used as a criterion measure for each manager was the mean of the scores received on the two forms of the scale. The rating scores assigned by the district managers were subjected to an analysis of variance to determine if there were significant differences between the mean ratings made by the various district managers. The resulting *F* value of 2.15 was not significant at the 5% level of confidence, indicating that the differences in mean ratings may be attributed to chance alone.

The effect of job experience on the rating score was checked by comparing the mean rating scores obtained on nine managers with three to six months' experience with scores of men who had managed for more than one year. In the absence of a significant difference between these groups, using the *t* test for small samples, it appears that within the time limits studied, variation in experience is not associated with the average rating score.

Combination of the Criterion Measures. It is first necessary to examine the comparability of the three criterion measures in terms of their respective means and standard deviations. These data, together with the reliability estimates for the three measures, are summarized in Table 1. These figures are based on data corrected for the district effects.

This table indicates that the three measures are not directly comparable in their present form because they do not have equal variances nor equal scale units.

Table 1
Summary Data of the Criterion Variates

Criterion	Reliability Estimate	Mean	S.D.
Total controllable costs	.96	68.1	3.3
Raw material costs	.86	37.5	1.9
Rating score	.88	5.5	0.7

The intercorrelations between the measures presented in Table 2 indicate the extent to which the criterion variates overlap. In interpreting this table it should be noted that controllable costs and raw material costs are experimentally dependent, and even the rating score may not be entirely independent of the above two because the rater's knowledge of a manager's operating data might influence some of the rater's responses to the rating form. The raters, however, did not see any operating data after district corrections had been made. The intercorrelations of Table 2 suggest that these variables may have a conspicuous common element which we may call managerial success or ability on the job. On the supposition that a common factor is being measured, it is possible to combine these three measures into a composite criterion. The individual measures on each variable were converted

Table 2
Intercorrelations of Criterion Variates

Comparison	r
Controllable costs vs. raw material costs	.65*
Controllable costs vs. rating score	.33
Raw material costs vs. rating score	.41

* This coefficient of .65 may be regarded only as an approximation of the true correlation because of two conditions: (1) raw material costs are actually a portion of total controllable costs and hence the two measures overlap, and (2) raw material costs and total controllable costs are both ratios which have the same denominator, viz., total sales. Correction for the former source of error may be made by Peters and Van Voorhis formula for the correlation between overlapping arrays (13, p. 215–217). An application of this correction here yields a value of .10 which seems unreasonably low because it is based upon the assumption that the second effect mentioned is negligible and that the two ratios have a zero correlation. This second effect can be checked by a partial correlation technique which is inappropriate here because of the small number of cases.

into normalized standard scores, using the percentile method, and the three standard scores were averaged for each individual manager. This procedure gave equal nominal weights to each of the three variables, but the effect of the intercorrelations and the variance relationship between total controllable costs and raw material costs actually gives a greater effective weight to raw material costs. Because of the greater importance attached to this latter variable by top management, the weighting used here is in the desired direction and appears to yield a satisfactory combined criterion measure.

Preliminary Test Battery

The preliminary test battery was assembled and administered to all bake shop managers. Those managers who had three or more months experience formed the criterion group which was used in the evaluation and item analyses of the several tests. Following is a brief description of these tests and preliminary results obtained from them.

General Mental Ability. The research of previous investigators (1, 6, 12) indicates that there is some positive relationship between scores on short mental ability tests and job success in certain managerial or supervisory positions. The Wonderlic Personnel Test, a 12 minute revision of the Otis S-A Test, scored by taking the number of correct responses made in the 12 minute time interval and correcting the score for age by the use of Wonderlic's table (19), was included in the present battery. The scores of the population of 79 managers ranged from 7 to 34 with a mean of 19.8 and an S. D. of 6.5. A significant relationship between tests score and educational level is revealed by a correlation of .39. The correlation of test score with the composite criterion was .13, while a correlation of .20 was obtained between test score and size of unit the manager operated. The latter value just fails to be significant, for an r of .22 is required for the 5% confidence level for this size of sample.

Preference, Interests and Attitudes. Previous investigators agree that the personality of the manager or supervisor is one of the most important single elements contributing to job success. Tests in this area have thus far been rather unsuccessful as selection instruments, partially because such paper and pencil inventories can often be "beaten" by the applicant. In the selection situation the applicant's motives may lead to responses which he thinks will help him obtain the position. A second shortcoming of the usual "personality test" is that it is scored in terms of a number of traits such as dominance, introversion, frankness, etc. Since it cannot be precisely determined if these traits are required for success on a given job, the industrial validation of such a "test" is a difficult and often impossible procedure.

Jurgensen (7) has recognized the shortcomings of the common types of personality inventories when used as selection instruments, and has constructed a device which he believes possesses distinct advantages in the industrial situation. He has utilized the forced choice technique of item arrangement which has also been used in the selection of army officers (15). The three principal advantages of Jurgensen's Classification Inventory are that: 1. the applicant is generally not able to predict the "right" answers when attempting to secure a job; 2. the test is scored and validated on a specific job and a scoring key developed for the job in a given company; and 3. no hypothetical "trait scores" are necessary because each item can be correlated separately with the criterion.

A tentative scoring key for bake shop managers was developed in accordance with the procedure recommended by Jurgensen (8). The criterion population of 79 managers was split into two groups which comprised the highest 27% and lowest 27% in terms of the composite criterion. The scoring keys were based on items which differentiated between the high and low criterion groups at the 10% level of confidence or better. The resulting scores of the managers correlated .64 with the composite criterion, but this value must be interpreted with extreme caution because 54% of this population was used in the construction of scoring keys for this test.

Job Information. In most instances, managers or supervisors are selected from existing employees and are expected to possess some knowledge of the production work or technical specialty performed by the persons they will supervise. Bake shop managers are expected to possess considerable baking experience and they must be able to recognize why products are below standard. In addition, it seemed desirable to know the amount of baking information possessed by a manager-candidate so that the training of the individual could be arranged accordingly. A test was constructed in which the majority of items were directed towards measuring the "diagnostic baking sense" of the manager. The following multiple choice item is an example:

Which one of the following conditions is most likely to result in dull crust color on Danish pastry? A. Not enough dusting flour; B. Underproofing; C. Low egg content; D. Old dough.

The Baking Knowledge Test consisted of 46 items. All 79 managers in the original population were tested and a correlation of .16 was obtained between test score and the composite criterion. The test was then subjected to an item analysis in order to determine which items differentiated between good and poor managers. The highest and lowest 27% of the criterion group were again used for the item analysis. When the Baking Knowledge Test was rescored, using only 12 items which met the criterion of discrimination, the scores correlated .37 with the composite criterion of managerial ability. This figure must be regarded as a spuriously high validity estimate because 54% of this population was used in item analysis of the Baking Knowledge Test.

Judgment in Managerial Problems. A number of items were constructed which represent some of the decisions and judgments a bake shop manager must make. The items were arranged in multiple choice form, but the respondent is required to select the first, second and third best choices from the alternatives in each item:

Assume your best selling item is a pecan ring that sells for 35¢ each. This item accounts for 10% of your sales. Pecans are selling for 40¢ per pound. Then the price of pecans jumps to $1.25 per pound because of crop failure. Would you—

    A. Stop making pecan rings and try to build up sales on another item;
    B. Increase the price of pecan rings to 60¢ because the material cost will be in line at this price;
    C. Leave the price the same and try to make up your loss by increasing prices a little on other items;
    D. Use only one-third as many pecans as before and leave the price the same.

Items of this type have been grouped together as the Federal Management Test. A rank order of several choices per item was required because Cardall (4) found the second or third choice responses to an item are sometimes more discriminating than the first or "best" choice. The item analysis and scoring key were based on the responses of the top and bottom 27% of the criterion group.

The scoring key was constructed by computing the percentage of "high" and "low" criterion managers responding to each item alternative as a first, second or third choice. This analysis yielded 12 items containing one or more responses which successfully differentiated between the high and low criterion groups. Scores on this test correlated .52 with the composite criterion.

A second test was constructed which was designed to measure managerial judgment in a specific context. One of the important daily duties of the manager is that of ordering the quantity and selection of baked goods to be produced for each day. This requires an accurate estimate of expected volume of business on the following day so that the shop will not be sold out before closing time and, conversely, that few or no items need be carried over as "stales." The Bake Order Problem was constructed on the assumption that a hypothetical store under given weather conditions could serve as a basis for measuring an individual's ability to use correct judgment in making out the order. After this problem was administered to the criterion population, responses to different portions of the problem were analyzed to determine if there were significant differences between the high and low 27% criterion groups in terms of quantity of each item ordered and variety of items ordered. A total of 13 analyses were made on different portions of the problem, but all results were negative and this problem was omitted from the revised battery.

Biographical Data. As early as 1922 Goldsmith (5) found that biographical information items were helpful in the selection of insurance salesmen. Uhrbrock and Richardson (18) and the Army (15) have both used such items in personnel selection batteries. Personal data were collected in the present study by means of a Biographical Information Blank which included 33 items. Since these items were first being used on present managers, it was necessary for the testee to answer each item as it applied when he first became a manager. For example:

How many of the following did you own or were you buying just before you became a Federal manager? (Mark as many answers as apply): A. Stocks or bonds ($100 to $300); B. Stocks or bonds (more than $300); C. A house; D. Home furnishings; E. A bakery; F. A car.

An item analysis was performed on the 33 biographical items using the high 27% and low 27% of the criterion groups responding to each alternative. The resulting response frequencies for most item alternatives were extremely small and consequently the validity estimates were unstable. It was also found that only a small number of items discriminated between the two criterion groups. For these reasons the Biographical Information Blank was omitted from further consideration.

Name and Number Checking. The typical bake shop manager spends about one hour per day on report work and simple bookkeeping. In the light of these activities the Minnesota Clerical Test was included in the preliminary battery. The results obtained indicated a lack of correlation between responses on this test and the composite criterion. The correlation was .19 between "numbers" score and the criterion and −.15 between "names" score and the criterion. The 79 managers obtained a mean score of 108.7 and an S. D. of 30.3 on the Numbers Test and a mean of 92.5 and S. D. of 28.5 on the Names Test. These mean scores are rather low when compared to norms for male clerical workers reported by Andrew and Paterson (2). It may be hypothesized that the managers in the present study do not perform enough clerical work or can do this work at their own pace in a manner which is not identical with the perceptual speed factor which probably operates in the Minnesota Clerical Test.

The Revised Battery

The following tests were included in the revised battery: Classification Inventory, Baking Knowledge Test, Federal Management Test and

Wonderlic Personnel Test. This battery was administered to a new population of 85 manager applicants already in the employ of the company as bakers. This population was used to establish tentative norms for the tests. Thirty-three of these applicants were selected for manager training and subsequently became managers. These men formed the cross validation population. In addition, 23 present managers who had taken the preliminary battery were retested on the revised battery to furnish reliability data on the various tests.

The population of 85 applicants had a mean age of 32.8 years, a mean education of 10.0 grades and a mean of 11.2 years civilian baking experience. Corresponding data for the original group of managers indicates a mean age of 42.7 and mean education of 10.0 grades.

Reliability of the Battery. Reliability data are based on 23 managers who were retested seven months after the original testing. The reliability estimates for the battery are presented in Table 3.

The low reliability of the Federal Management Test casts doubt on its usefulness.

Table 3
Test-Retest Reliability Estimates for Tests in the Revised Battery (N = 23)

Test	Reliability (r)
Classification Inventory	.78
Baking Knowledge	.77
Federal Management	.46
Wonderlic Personnel Test (Forms A and B)	.85

The mean scores obtained by these 23 men on test and retest sessions were analyzed to determine if any significant shifts had occurred. It was found that there was no significant change in group mean scores for the Classification Inventory and Federal Management Test. However, mean scores on both the Baking Knowledge Test and Personnel Test increased significantly. These differences were significant at better than the 1% confidence level. It is possible that familiarity with the testing situation, positive practice effect and memory may have accounted for the increase in scores on the latter two tests.

Intercorrelations of Test Scores. One of the interesting findings obtained from the applicant group was a revised set of intercorrelations of the several tests. The test intercorrelations from the criterion population and from the applicant population are presented in Table 4. This table shows that the original intercorrelations on all tests except the Personnel Test were much higher than the values obtained from the applicant

population. This finding might be anticipated because the Personnel Test was the only one of the four tests which was not item analyzed or scored on the basis of the responses of the original population of managers. None of the intercorrelations involving the Personnel Test shifted markedly on the new population, whereas the other values show a definite decrease for the applicant group. Such results accent the fact that correlational values obtained from a population which is used in the construction or item analysis of tests will generally be spuriously high.

Table 4

Intercorrelation of Tests

Note: Superior values in each cell are based on 85 applicants.
Values in parentheses are based on criterion population.

	Baking Knowledge	Federal Management	Personnel Test
Classification Inventory	−.16 (.44)	−.02 (.48)	.38 (.28)
Baking Knowledge		.04 (.41)	.17 (.19)
Federal Management			−.10 (.00)

The only significantly positive intercorrelation for the applicant population is between the Personnel Test and the Classification Inventory. Thus these two measures are somewhat dependent, although the size of the correlation does not indicate a marked "overlap." Only the Baking Knowledge Test correlated significantly (r = .45) with number of years of baking experience. None of the tests correlated significantly with age and only the Personnel Test correlated significantly with number of years of education (r = .43).

Validity Data. Thirty three of the 85 applicants were appointed as unit managers. These men were given a Company manager training program before being assigned to a managerial position. The 33 men had actually been managing for an average of 8.8 months when the follow-up study was conducted and criteria of their job performance were collected. Criterion data were obtained on these men by the same procedures as were used in the original study and the composite criterion used in the follow-up study is identical in composition with that used in the original study.

It should be pointed out that the district manager (who was the rater) knew the new manager's test scores in ten out of the 33 cases. This

possibility of criterion contamination is probably slight, especially because the ratings constitute only a portion of the composite criterion.

The validity coefficients of the various tests in the revised battery are presented in Table 5. Correlation coefficients are not very appropriate measures for such a small sample. In the present sample the Classification Inventory has a coefficient which is significant at better than the 5% level. None of the other coefficients in Table 5 are significant.

An alternate method of estimating the validity of the several tests is to compare the test scores of men who had high criterion scores with the scores of men who received low criterion scores. For this analysis the population of 33 managers was divided into the best 16 and poorest 16 on the basis of the composite criterion measures. The remaining case was omitted from this analysis. The "high" group of managers made significantly higher scores than the "low" group on two of the tests—the Classification Inventory and the Personnel Test. In both cases, the t values for the differences between mean scores were significant at better than the 5% confidence level. The Baking Knowledge Test and the Federal Management Test failed to differentiate between good and poor managers.

Table 5
Correlations Between Test Scores and Criterion (N = 33)

Test	Validity Coefficient
Classification Inventory	.39
Baking Knowledge	−.12
Federal Management	.06
Wonderlic Personnel Test	.26

It was decided to eliminate the Federal Management Test from further use as a selection device because of its lack of validity and its low reliability coefficient ($r = .46$). Although the Baking Knowledge Test did not appear to be a valid predictor of managerial success, this test was retained in the battery because it would be useful in determining the amount of baking training the applicant would require before he was installed as a manager.

The scattergrams of test scores against criterion measures for the Classification Inventory and Personnel Test were inspected in an attempt to set cutting scores for these tests. It was found that a cutting score of 16 on the Personnel Test would have eliminated 44% of the "low" managers (poorest half on the criterion) but would have eliminated only 6% of the "good" managers. This raw score of 16 is equivalent to a percentile rank of 25, as determined from the population of 85 applicants.

Similarly, a passing score of 25 (percentile rank of 48) on the Classification Inventory would have eliminated 50% of the "poor" managers and 36% of the "good" managers. If these two tests were used together and both of the above cutting scores had been used in combination, 63% of the "poor" managers would have been rejected as opposed to 36% of the "good" managers. On the basis of these combined cutting scores, 45% of the 85 applicants would have been considered as acceptable for managerial training. However, these cutting scores should be considered as tentative until they are validated on a second independent sample.

Summary

A study has been made of the prediction of managerial success in a retail-manufacturing bakery chain. An empirical approach has been used to select tests and to select and weight items on the basis of the responses of a criterion population of 79 managers. Three criterion measures were obtained on these managers and these were combined into a composite criterion score for each manager.

A preliminary battery of seven tests was administered to the 79 managers. A subsequent item analysis suggested that a revised battery be assembled. This contained the Baking Knowledge Test, the Wonderlic Personnel Test, the Classification Inventory and the Federal Management Test. This revised battery was administered to 85 applicants for managerial positions. New test norms, intercorrelational data and reliability data were obtained from this population. All tests in the revised battery except the Federal Management Test had acceptable reliability coefficients.

The validity of the battery was estimated by comparing the test scores and criterion scores of 33 of the applicants who subsequently became managers. Only one of the tests—the Classification Inventory—had a validity coefficient which was significantly different from a zero correlation at the 5% level of confidence. A comparison of the mean test scores of the upper and lower halves of this group, based on the criterion, indicate that both the Wonderlic Personnel Test and the Classification Inventory significantly differentiated between the good and poor managers. The Federal Management Test and the Baking Knowledge Test lacked validity, but the latter test was of value in determining how much additional baking training was required by the individual.

References

1. Achard, F. H., and Clarke, F. H. You can measure the probability of success as a supervisor. *Personnel*, 1945, **21**, 353–373.

2. Andrew, D. M., and Paterson, D. G. *Manual for the Minnesota Clerical Test.* New York, The Psychological Corporation, 1946.

3. Beckman, R. O., and Levine, M. Selecting executives: an evaluation of three tests. *Person. J.*, 1930, 8, 415–420.

4. Cardall, A. J. *Manual for the test of practical judgment.* Chicago, Science Research Associates, 1942.

5. Goldsmith, D. B. The use of the personal history blank as a salesmanship test. *J. appl. Psychol.*, 1922, 6, 149–155.

6. Harrell, W. Testing cotton mill supervisors. *J. appl. Psychol.*, 1940, 24, 31–35.

7. Jurgenson, C. E. Report on the "Classification Inventory," a personality test for industrial use. *J. appl. Psychol.*, 1944, 28, 445–460.

8. Jurgensen, C. E. *Manual for Classification Inventory*, privately printed, 1947.

9. Knauft, E. B. A classification and evaluation of personnel merit rating methods. *J. appl. Psychol.*, 1947, 31, 617–625.

10. Knauft, E. B. Construction and use of weighted check-list rating scales for two industrial situations. *J. appl. Psychol.*, 1948, 32, 63–70.

11. Mandell, M. Testing for administrative and supervisory positions. *Educ. & Psychol. Meas.*, 1945, 5, 217–228.

12. Mandell, M., and Adkins, D. C. The validity of written tests in selection of administrative personnel. *Ed. & Psychol. Meas.*, 1946, 6, 293–312.

13. Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases.* New York, McGraw-Hill, 1940.

14. Richardson, M. W., and Kuder, G. F. Making a rating scale that measures. *Person. J.*, 1933, 12, 36–40.

15. Staff, Personnel Research Section, Adjutant General's Office. The forced choice technique and rating scales. *Amer. Psychologist*, 1946, 1, 267 (Abstract).

16. Thompson, C. E. Selecting executives by psychological tests. *Ed. & Psychol. Meas.*, 1947, 7, 773–778.

17. Thurstone, L. L. Attitudes can be measured. *Amer. J. Sociol.*, 1928, 33, 529–554.

18. Uhrbrock, R. S., and Richardson, M. W. Item analysis: the basis for constructing a test of supervisory ability. *Person. J.*, 1933, 12, 141–154.

19. Wonderlic, E. F. *Wonderlic Personnel Test Manual.* Privately printed, 1945.

A Note on Mechanical Aptitude of West Texans

Albert Barnett

Texas Technological College

Two tests claiming to measure mechanical aptitude have been rather widely used at Texas Technological College. For a number of years, freshmen, as part of an orientation program, were given the Revised Minnesota Paper Form Board, a paper-and-pencil test requiring the testee to combine in his imagination a few disarranged geometrical plane figures to form one large figure and select the correct answer from among four or five suggested solutions. The Minnesota Spatial Relations test requires that the testee fill a number of irregular holes in each of four form-boards with the appropriate cut-out blocks, no two of which are alike, the score being the number of seconds required to complete the task. This test has been used for some time on an individual basis at the Texas Tech. Guidance Center.

It is evident that neither of these tests places much, if any, emphasis on mere hand skills, but on the mental factor of space relationship, which, it is claimed, accounts in part for mechanical aptitude.

During the fall semester of 1941, the Revised Minnesota Paper Form Board AA, was run on 371 freshmen (mainly from West Texas) of the Arts and Science Division of Texas Technological College. Their mean chronological age was between 18 and 19 years, the approximate range being 15–23 years. Their median score was 42.5 equivalent to the 70%ile on the norms of liberal arts freshman men, whose median score was 38. The fact that these young freshman liberal arts boys, on the average, excelled 70 per cent of the standardization group in mechanical aptitude was merely noted, but not explained.

During several months in 1947–1948, a record was made of the scores of men on the Minnesota Spatial Relations Test coming to the Texas. Tech. Guidance Center for vocational advisement. These men, mainly in their twenties and thirties, came from several different West Texas counties, and represented every educational level from the illiterate to the college graduate. Each man was tested individually by a trained psychometrist. Results are shown in Table 1. It may be noted that the mean time required by this sample of 383 men to complete the test was 973.5 seconds, which compares to the mean (apparently) of 1279 seconds for the norms furnished by the publishers of the test. The difference be-

Table 1

Minnesota Spatial Relations Test Results
(383 Men, Texas Tech. Guidance Center)

Seconds Required for Completing Test	f	Percentile	Standard Score
500–599	3	99.19	6.90
600–699	16	96.72	6.45
700–799	59	86.97	6.00
800–899	87	67.99	5.55
900–999	74	47.06	5.11
1000–1099	63	29.25	4.66
1100–1199	29	17.29	4.21
1200–1299	21	10.79	3.76
1300–1399	11	6.63	3.31
1400–1499	6	4.42	2.86
1500–1599	10	2.34	2.41
1600–1699		1.04	1.96
1700–1799	1	.91	1.52
1800–1899	1	.65	1.07
1900–1999	2	.26	0.62
	383		

$$M = 973.5 \qquad \epsilon_M = 11.4 \qquad \sigma = 222.31$$

Table 2

Norms for Men Compared to Achievement at the Texas Tech. Guidance Center

Letter Rating	Mid-Sigma Score	Time in Seconds for all Four Boards	
		Texas Group	Published Norms
A	7.0−	Up to 639	Up to 936
B	6.0	640–861	937–1131
C	5.0	862–1085	1132–1427
D	4.0	1086–1307	1428–1934
E	3.0−	1308 and above	1935 and above

tween the Texas group and the norm group is revealed in Table 2, which shows the score range equivalent to the letter ratings on the test for the Texas group compared to the published test norms.[1]

As yet, no satisfactory explanation has been found for this superiority (as tested) of West Texas men in mechanical aptitude. It is true that the region from which the Texas group came is one of mechanized

[1] *Minnesota Spatial Relations Test: Examiner's Manual* (Minneapolis: Educational Test Bureau), p. 3.

farming.   Most of these men had been accustomed to tractors and other
machines since boyhood.   Some worked with machines in the oil fields of
this region.   The Spatial Relations Test, however, is supposed to be, as
stated in the *Examiner's Manual,* "relatively free from the influence of
previous mechanical experience."

As stated previously, the men who were tested at the Texas Tech.
Guidance Center were young men.   It is not known whether or not they
were as a group younger than the standardization group.   As a check
on the possible influence of age on test performance, the Texas group
was separated into two discrete sub-groups, namely; those requiring
one thousand seconds or more to complete all four boards of the test and
those who completed the test in eight hundred seconds or less.   The
former group had a mean age of 26.7 years as compared to 24.2 for the
latter, the standard error of the difference being .66 and the critical ratio
3.76.   While it is true that the poorer performance is associated with the
older group, there is much over-lapping.   Furthermore, it is possible that
among the older men, those who had failed to adjust occupationally be-
cause of poor mechanical aptitude, tended to present themselves for
testing and guidance more than was the case of those who had adjusted.
Further study needs to be made on the relationship of tested hand skills
to tested mechanical aptitudes.

Work Satisfaction and Work Efficiency of Vocational Counselors as Related to Measured Interests *

Salvatore G. DiMichael

Office of Vocational Rehabilitation, Federal Security Agency

This article reports another phase of a broad study designed to obtain a more complete understanding of personnel engaged as vocational rehabilitation counselors for the civilian disabled. A previous article described the experimental study which devoted major attention to a determination of the pattern of measured interests and of the relationships between measured and self-estimated interests for a group of counselors. It was found that the typical profile of measured interests on the Kuder Preference Record was sharply differentiated from the general population; that the highest median vocational interest areas were Social Service (98 %'ile), Persuasive (82 %'ile), and Literary (65 %'ile); that the reliability coefficients for the scales ranged from .70 to .89 with an average time interval of 5 months between tests; that self-estimated interests generally correlated to a substantial degree (median $r = .56$) with measured interests; and that when the counselors had previous knowledge of their Kuder results, it did not change the subjective expressions of their interests in the direction of the objective preference scores (1).

In the present report, the experimental investigation primarily deals with the possible relationships between the measured interests of vocational rehabilitation counselors and their work satisfaction, and work efficiency. This study sought to determine whether the Kuder results could give a basis for predicting varying degrees of work satisfaction and of work efficiency among a selected population of counselors who were already on the job when the experiment was begun.

On the basis of a critical review of the experimental literature on the Kuder Preference Record, Super states that "the evidence justifies the conclusion that the Kuder Preference Record has now been sufficiently well standardized and validated for use in vocational guidance. . . . More research needs to be done before the Record can be considered a well-understood instrument, but it is already a valuable tool in the counselor's kit" (6, p. 191).

Evidence of the validity of the Kuder Preference Record in terms of enjoyment and efficiency on the job is, according to Kuder's 1946 manual

(4), found in only one study. In the latter, Hahn and Williams (3) reported significant differences between mean scores on the clerical scale for satisfied and dissatisfied workers of three clerical groups of women Reservists in the Marine Corps.

Method

While conducting orientation institutes for counselors engaged in the State-Federal vocational rehabilitation program for physically and mentally disabled civilians, the author administered the Kuder Preference Record to the trainees. They were assured of the confidentiality of individual results and were requested to turn in their interest profiles for use in an experimental investigation about the interest patterns of rehabilitation counselors. Five months later on the average, they were requested to retake the Kuder and also to fill out a Survey Sheet which recorded their degree of interest with the job of counseling taken as a whole and with distinguishable phases of it. At the same time, a prepared Job Rating Schedule was sent to each of the counselor's supervisors who were requested to rate the men for efficiency on the job as a whole and various phases of it. Of the initial group of 134 counselors, 10 had resigned in the meantime and 24 had some of the necessary records missing; the remaining number of 100 is referred to collectively as Group A.

Group B was made up of 46 counselors who had never taken the Kuder inventory before. They first were requested to fill out the items in the Survey Sheet which included questions about work satisfaction. Then the Preference Record was administered. In the present study, data on Group B enter into the experimental results only on job satisfaction. No job efficiency ratings were secured on this group.

Counselors' Ratings on Work Satisfaction

The counselors were asked on the Survey Sheet (graphic rating scale method) to rate the degree of their liking for the job as a whole and for particular phases of the job (9 items).

The checked ratings were converted to numerical scores from 0 to 20. The results in terms of means and standard deviations for Groups A and B are listed in Table 1. By a comparison of the averages of the counselors' ratings on the different scales, it is possible to estimate the relative degrees of satisfaction with several phases of their work. The highest amount of work satisfaction seemed to be derived from "interviewing clients" and the "job as a whole." Other phases of the work which gave high average satisfaction scores were "promoting the program to the public," "contacting employers for jobs," "reading scientific

Table 1

Self-Ratings of Vocational Rehabilitation Counselors on Degree of Work Satisfaction with Their Job as a Whole, and with Particular Aspects of Their Job

Work Satisfaction Scale*	Groups**	Mean Rating	S.D.	"t" ratio†
Whole Job	A	18.8	1.85	2.05
	B	17.9	2.60	
Interviewing	A	19.0	1.55	1.18
	B	18.6	1.97	
Promoting the Program	A	16.4	3.79	1.50
	B	15.2	4.55	
Contacting Employers for Jobs	A	15.9	3.48	.38
	B	15.7	3.66	
Reading Scientific Lit. on Rehabilitation	A	16.0	3.31	.77
	B	15.5	3.49	
Experimenting with Guidance Techniques	A	16.0	3.37	− .12
	B	16.0	3.16	
Writing Case Histories	A	12.7	4.04	− .28
	B	12.9	5.27	
Handling Clerical Details	A	10.5	4.86	1.80
	B	8.9	5.07	
Rehabilitation Work After Business Hours	A	12.5	4.50	3.38
	B	9.4	5.29	

* Conversion of ratings to numerical scores was made on basis of a scale of units from 0 to 20.

** N for A = 100; for B = 46.

† Values required for statistical significance at 5% level of confidence = 1.98; at 1% of level of confidence = 2.61 (5, pp. 212–3).

literature on rehabilitation," and "experimenting with guidance techniques." Less enjoyment was reported from "writing case histories" and doing "rehabilitation work after hours." The phase of the job liked least was "handling clerical details."

From an inspection of the frequency distributions on the above items, it appeared that the converted scores on the different rating scales were not normally distributed. All but one appeared to be considerably skewed. The range of scores was very wide on all items except two, namely "whole job" and "interviewing." In the latter, the ranges were highly restricted to the upper levels of the scales.

On each of the items dealing with work enjoyment, the differences in average self-ratings between Groups A and B appeared slight. How-

ever, it was necessary to test the differences statistically in order to be able to state definitely that they were or were not due to chance. Accordingly, the "t" ratios were computed and the only significant differences between the groups related to the items "enjoy the job as a whole" and "enjoy rehabilitation work after business hours." These results signify that Group A claimed a higher degree of satisfaction than Group B in the job as a whole and in overtime work, and also show that the differences in average self-ratings on each of the other items are too small to be regarded as statistically significant.

Work Satisfaction and Measured Interests

In setting up the study, one important hypothesis to be investigated was that certain Kuder Preferences results could be used to predict greater enjoyment with various phases of the total job, as well as with the job as a whole. Thus, it seemed logical to expect that persons with higher scores in the Kuder scales which distinguished the counselors from the general population, namely Social Service, Persuasive, and Literary, probably would be more satisfied with the job of counseling as a whole. It also seemed logical to expect that counselors who came out higher on the Scientific scale of the Kuder would experience more job interest and satisfaction in experimenting with guidance techniques; and that counselors higher in the Literary scale of the Kuder would be more apt to enjoy writing up case histories; and that counselors who scored higher in the clerical scale of the Kuder would be less annoyed with the handling of the clerical details.

The possible relationships between Kuder scores and work satisfaction were determined by computing correlation coefficients between variables that logically could be suspected of showing a significant degree of relationship.

Because the scores on job satisfaction did not appear to be distributed normally, as noted above, it was necessary to consider the possibilities of curvilinear relationships between scores on the Preference Record and the job satisfaction scales. Parenthetically, it may be mentioned that the scores on the Preference scales appeared to be normally distributed with the exception of the Artistic scale, in which the scores seemed to be skewed positively. A scatter diagram was prepared for each of the paired variables shown in Table 2 for Group A. An inspection of each of the diagrams and of the empirical regression lines for the prediction of job-satisfaction scores from Kuder scores indicated no curvilinearity. A similar analysis on Group B, with a smaller number of cases than Group A, did not appear to be warranted. The statistical data are

presented in Table 2. It may be seen that six correlation coefficients are statistically significant.

The two paired variables which showed a statistically significant correlation for *both* groups were enjoyment in "contacting employers for jobs" and the Kuder Persuasive scale. The evidence was not as clean-

Table 2
Relationship Between Kuder Preference Scores and Job Satisfaction as a Vocational Rehabilitation Counselor

Satisfaction Scales	vs.	Kuder Preference Scales	r	Group
Job as a Whole		Pers.	.16	A*
Job as a Whole		Pers.	.12	B**
Job as a Whole		Soc. Ser.	.13	A
Job as a Whole		Soc. Ser.	.29	B
Interviewing Clients		Pers.	.15	A
Interviewing Clients		Pers.	.00	B
Interviewing Clients		Soc. Ser.	.06	A
Interviewing Clients		Soc. Ser.	.43	B
Promoting the Program		Pers.	.17	A
Promoting the Program		Pers.	.13	B
Contacting Employers for Jobs		Pers.	.28	A
Contacting Employers for Jobs		Pers.	.30	B
Reading Scientific Literature on Rehab.		Sci.	−.02	A
Reading Scientific Literature on Rehab.		Sci.	.01	B
Reading Scientific Literature on Rehab.		Lit.	.05	A
Reading Scientific Literature on Rehab.		Lit.	.04	B
Experimenting with Guidance Techniques		Sci.	−.01	A
Experimenting with Guidance Techniques		Sci.	−.03	B
Writing Case Histories		Sci.	−.09	A
Writing Case Histories		Sci.	−.25	B
Writing Case Histories		Lit.	.03	A
Writing Case Histories		Lit.	.15	B
Writing Case Histories		Soc. Ser.	.00	A
Writing Case Histories		Soc. Ser.	−.08	B
Handling Clerical Details		Cler.	.32	A
Handling Clerical Details		Cler.	.21	B
Rehabilitation Work After Hours		Pers.	.12	A
Rehabilitation Work After Hours		Pers.	.30	B
Rehabilitation Work After Hours		Soc. Ser.	.00	A
Rehabilitation Work After Hours		Soc. Ser.	.05	B

* ($N = 100$.) Values of correlation coefficients required for statistical significance are .197 at the 5 per cent level of confidence and .256 at the 1 per cent level of confidence (5, p. 212).

** ($N = 46$.) Values required for statistical significance are .291 at the 5 per cent level of confidence and .376 at the 1 per cent level of confidence (5, p. 212).

out for the other paired variables which showed a statistically significant correlation coefficient for one group but not the other. The variables, enjoyment in "interviewing clients" and the Kuder Social Service scale, showed a correlation coefficient for Group A which was very close to zero, although for Group B, the same variables showed a significant relationship beyond zero. The latter difference is difficult to explain satisfactorily.

Work Efficiency and Measured Interests

Another important phase of this study was to determine the relationships of the Kuder interest scores to the supervisory ratings on job efficiency. The results should indicate the possible value of the interest scores in predicting successful performance in the job, or in particular phases of the job. For example, did the high interest scores in the Persuasive, Literary, and Social Service scales have a relationship to proficiency on the job as a whole, and on such phases of the job as interviewing clients, interpreting psychological tests, using community resources, having high production and doing quality counseling? Similarly, did high Scientific interest scores correlate with effectiveness in experimenting on and trying out new professional techniques, etc.

At present, there are no satisfactory objective devices to evaluate job efficiency in vocational rehabilitation counseling. For this reason, the graphic rating-scale method was used, accompanied by instructions which sought to improve the reliability and validity of the ratings.

The items rated were: a. counseling efficiency as a whole; b. conducting counseling interviews; c. interpreting psychological test results to the client; d. effective use of community resources; e. imparting occupational information to the client; f. writing up case reports; g. handling financial records for client's rehabilitation expenses, making up the flow sheets, keeping field sheets up to date; h. making talks, speeches and promoting the program to the general public; i. making contacts with employers to secure job opportunities for his clients; j. reading current scientific articles, books, and reports on rehabilitation topics; k. experimenting on and trying out new techniques of counseling and guidance; l. continued work after regular hours; m. production record on rehabilitations; and n. quality of work.

The frequency distributions of the efficiency ratings as converted into numerical scores from 0 to 20 were inspected for indications of normality. The distributions generally appeared to be very peaked at the center of the scale, usually with shorter peaks at the guide points designated as "passable" and "very good" on the graphic scales. The scores on each item spread over almost the entire range, and did not

appear markedly skewed. These indications made it necessary to consider the possibility that the relationships between job-efficiency and Kuder scores might be curvilinear. Accordingly, scatter diagrams were prepared, as well as empirical regression lines for the prediction of efficiency ratings from Preference scores. Inspection of the regression lines indicated no curvilinear relationships between the job-efficiency and the first Kuder scores. There seemed to be no reason to assume that the

Table 3

Relationship Between Kuder Preference Scores and Job Efficiency of Vocational Rehabilitation Counselors as Rated by Supervisors

Job Efficiency vs. Kuder Scale	on 2nd Kuder	on 1st Kuder
Whole Job vs. Mech.*	−.12	−.04
Whole Job vs. Comp.	.08	.01
Whole Job vs. Sci.	.02	−.03
Whole Job vs. Pers.	.01	.10
Whole Job vs. Art.	−.09	−.14
Whole Job vs. Lit.	.06	.03
Whole Job vs. Mus.	.07	.02
Whole Job vs. Soc. Ser.	−.04	−.02
Whole Job vs. Cler.	.13	.07
Interviewing vs. Pers.	.14	.16
Interviewing vs. Soc. Ser.	.09	.00
Interpreting Tests vs. Sci.	.04	.08
Use of Community Resources vs. Soc. Ser.	.05	.13
Imparting Occupational Information vs. Mech.	.13	.14
Imparting Occupational Information vs. Comp.	.06	.07
Imparting Occupational Information vs. Sci.	.14	.03
Imparting Occupational Information vs. Pers.	.05	.17
Imparting Occupational Information vs. Soc. Ser.	−.09	−.07
Imparting Occupational Information vs. Cler.	.00	.00
Writing Case Histories vs. Sci.	−.03	−.03
Writing Case Histories vs. Lit.	.07	.10
Writing Case Histories vs. Soc. Ser.	−.05	−.07
Handling Records vs. Cler.	.05	.08
Publicly Promoting the Program vs. Pers.	.24	.32
Contacting Employers vs. Pers.	.11	.19
Reading Scientific Literature vs. Sci.	−.07	−.11
Reading Scientific Literature vs. Lit.	.26	.26
Experimenting with Guidance Techniques vs. Sci.	.04	−.02
Rehabilitation After Hours vs. Soc. Ser.	−.09	.10
Production Record vs. Pers.	−.10	.04
Production Record vs. Soc. Ser.	−.10	−.03
Quality of Work vs. Pers.	.01	.11
Quality of Work vs. Soc. Ser.	−.01	.05

* Values required for statistical significance at 5% level of confidence = .197; at 1% level of confidence = .256 (5, p. 212).

relationships would be of a different type between the job-efficiency and the 2nd Kuder scores.

The product-moment r's found between the Preference Record scores, both first and second tests, and the supervisory ratings of job efficiency are presented in Table 3. It will be seen that correlation coefficients were computed only between those pairs of variables which might be suspected of yielding statistically significant relationships. Of the 66 coefficients, only five were statistically significant.

These results show that higher Kuder scores on the Persuasive scale tend slightly but definitely to indicate better job performance in promoting the program to the general public, and that higher scores on the Literary scale tend slightly but definitely to indicate greater activity in keeping up with the scientific literature in the field of rehabilitation. The evidence is not as clean-cut for the statement that there is a real relationship between job efficiency in contacting employers for jobs for handicapped clients and higher scores on the Persuasive scale of the Preference Record. The latter variables are found to be related at the five per cent level of confidence when the first Kuder test score is considered, but there is no statistical significance when the second Kuder score is involved.

Supervisors' Ratings on Job Efficiency Elements

It is interesting to study the distributions of the supervisory ratings on the several scales. A comparison of the averages of the efficiency ratings on the different items may indicate that supervisors are more satisfied with counselors' performance in some respects than with others. Perhaps the differences in average scores roughly indicate relative strengths and weaknesses in counselors' performance in civilian rehabilitation counseling at least as regarded by the supervisors. The foreword which accompanied the efficiency rating forms instructed the supervisors to rate their counselors so that the middle of the scale would be approximately the average for all counselors. Upon this background of instructions, the ratings on the scales resulted in the scores presented in Table 4.

According to the average ratings of the supervisors, they were relatively well satisfied with the "quality of work" done by the rehabilitation counselors. This item ranked first in order of magnitude. The supervisors also seemed to be relatively pleased with the counselors' efforts in the aspects of the job having to do with community contacts because the items next in order of magnitude were "use of community resources," and "contacting employers for jobs." The "production record" was less satisfactory than the "quality of work." The counselors were rated lowest in the items, "experimenting on and trying out new professional techniques" and "reading scientific publications on rehabilitation topics."

They also were more unsatisfactory on "promoting the program to the general public" and "interpreting psychological tests." Although the statistical data are not presented in this article, it has been found that the differences between means having rank orders (5) and above as shown in Table 4, as compared with means having rank orders (10) and below are statistically significant beyond the 1 per cent level of confidence. This signifies that in a similar sample the differences between the more extreme means will appear again if the experiment were to be tried over again under the same conditions.

Table 4

Supervisory Ratings on Counselors' Job Performance in Civilian Rehabilitation Work

Item Rated	Mean*	S.D.
Quality of Work	12.6	3.46
Using Community Resources	12.2	3.89
Contacting Employers for Jobs	12.1	4.29
Interviewing Clients	12.0	3.52
Writing Case Histories	12.0	3.72
Job as a Whole	12.0	3.50
Handling Records	11.6	3.60
Production Record	11.5	4.60
Imparting Occupational Information	11.3	3.86
Interpreting Tests	11.0	3.47
Rehabilitation Work After Hours	10.5	4.12
Promoting Program to Public	10.5	4.14
Reading Scientific Publications on Rehabilitation	10.4	3.10
Experimenting with Guidance Techniques	10.0	3.62

* The ratings were converted into numerical scores from 0 to 20.

An analysis of the magnitude of the standard deviations for all the items makes interesting material for a further observation. The highest standard deviations appeared for the items, "production record," "contacting employers for jobs," "promoting the program to the public," and doing "rehabilitation work after hours." The lowest standard deviations appeared for the items, "reading scientific publications on rehabilitation," "quality of work," "job as a whole," and "interviewing clients." Two reasons may be ascribed for these differences in the magnitude of the standard scatter of scores. One is that the group of the highest standard deviations relates to items more easily counted numerically, and that the second group above relates to more difficult qualitative judgments. In other words, supervisors spread out their ratings on counselors' job efficiency more on "production record" rather than "quality of work" because the former is more objective. A second possible reason for the

higher and lower magnitudes of the standard deviations is that the counselors are more alike in the group of items including "quality of work" than in the group of items including "production record." The first reason is the preferred explanation for the differences in standard deviations.

Summary

Vocational Rehabilitation counselors were requested to take the Kuder Preference Record, to fill out a Survey Sheet which indicated their work satisfaction on the job, and also were rated for work efficiency by their supervisors. It was found that:

1. The counselors derived a high degree of satisfaction from their job as a whole, and from such phases of it as interviewing clients, contacting employers for jobs, promoting the program to the public, experimenting with guidance techniques and reading scientific literature. They least enjoyed the handling of clerical details, overtime work, and the writing of case histories.

2. Higher scores on particular Kuder scales had low but significant relationships to work satisfaction for only several aspects of the counselor's job. However, the magnitude of the correlations was much too low for purposes of individual prediction. Higher scores on the Kuder Persuasive scale indicated greater enjoyment in contacting employers for jobs. Other evidences of significant relationship were not as consistent, appearing for one experimental group but not the other.

3. Higher scores on particular Kuder scales had low but significant relationships to work efficiency for only several aspects of the counselor's job. However, the correlation coefficients were too low for purposes of individual prediction. Higher scores on the Kuder Persuasive scale indicated greater work efficiency in promoting the program to the public; higher scores on the Kuder Literary scale indicated greater efficiency in keeping up with literature on rehabilitation.

4. The efficiency of counselors was rated more alike in such job elements as quality of work, interviewing clients, keeping abreast of modern scientific literature, and in the job as a whole. The counselors were rated less alike in such aspects of job efficiency as production record, contacting employers for jobs, promoting the program to the public, and working after hours. These differences probably were due to greater difficulty in rating the counselors on items depending upon qualitative rather than quantitative judgments of job performance.

References

1. DiMichael, S. G. The professed and measured interests of vocational rehabilitation counselors. *Educ. Psychol. Measmt.* 1949, 9: 59–72.

2. Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill Book Co., Inc., 1942.

3. Hahn, M. E. and Williams, C. T. The measured interests of marine corps women reservists. *J. appl. Psychol.,* 1945, 29: 198–211.

4. Kuder, G. F. *Revised Manual, Kuder Preference Record.* Chicago: Science Research Associates, 1946.

5. Lindquist, E. F. *Statistical analysis in educational research.* Boston: Houghton Mifflin Co., 1940.

6. Super, D. E. The Kuder Preference Record in vocational diagnosis. *J. consult. Psychol.,* 1947, 11: 184–194.

Certain Rorschach Response Categories and Mental Abilities

J. R. Wittenborn

Yale University

It is common practice among Rorschach technicians to include, as a part of their personality appraisals, some remarks concerning the subjects' "intelligence," "intellectual potential," "mental capacity," or "intellectual efficiency." Such evaluations are derived by the examiners from a variety of considerations.

The Rorschach scoring categories most commonly used in estimating mental capacity or achievements are: a. the total number of responses (R); b. the number of whole responses, i.e., responses based on the whole card (W); c. the number of responses in which Human Movement is seen (M); and d. the form level of responses, i.e., the accuracy and detail with which forms are seen.

These aspects of a Rorschach record are not employed independently in making appraisals. For example, the number of whole responses is dependent upon the accuracy with which forms are perceived. Since ability estimates offered by Rorschach workers make use of a wide variety of informal cues, it must be emphasized that it is *not* a purpose of the present investigation to determine the nature of the relationship between mental test scores and a mental ability estimate based upon a *total* Rorschach evaluation.

The purpose of the present investigation is to examine the ability implications of certain objectively determined, quantitatively expressed classes of response which are unique to the Rorschach. Specifically the investigation is concerned with the *location* (i.e., the portion of the blot employed) and the *determinant* (i.e., the shading, color or projected movement employed in forming a response) factors; these are unique to the Rorschach. The content of perceptions, as well as their accuracy (form level), are general factors in projection and their significance is not unique to the Rorschach. Therefore, the content and form level of responses are not included in the present analysis.

In meeting the purposes of the experiment, the analysis of data is conducted with respect to the following questions:

1. What is the order of the relationships between certain Rorschach response scoring categories and test evidence of mental ability? Are they negligible relationships which permit the kind of gross distinctions

between ability levels that can be made from casual observations, or do they provide refined distinctions comparable to those provided by mental tests? If the relationships are of high order, it should be generally known so that they can be put to extensive use. Since Rorschach responses appear to be less a function of specific education experiences than the performances currently sampled by most mental tests, it is conceivable that the demonstration of high relationships could influence future mental test procedures.

2. What is the nature of the relationship between various classes of mental ability and various Rorschach response categories? Intelligence, general ability, etc., are words for groups of human abilities, but the groups have no standard consistency. If a full appreciation of a relationship between a Rorschach response category and a mental ability is to be had, the nature of the mental ability in question must be specified. Accordingly, in the present analysis measures of verbal, spatial, and numerical abilities are employed. In addition, general measures of scholastic ability are included. If a pattern of relationship could be demonstrated between certain Rorschach response categories and certain classes of mental ability, an improved understanding of both Rorschach responses and of the mental abilities in question might result.

The Experimental Plan

The subjects were a heterogeneous group of 68 Yale students who had been in a speeded reading course or had consulted the writer. The Rorschach tests used in this analysis were administered and scored by Klopfer trained examiners.

The ability data employed in the analysis, the results of the College Entrance Examinations and the results of the Yale Freshman Aptitude tests, were taken from the files of the Yale University Student Appointment Bureau. Scores for the following variables were taken from each student's file and used in the analysis:

I. Scholastic Ability: 1. First Semester Freshman Year grade average; and 2. General Scholastic Prediction for Freshman Year.

II. Verbal Ability: 1. College Entrance Scholastic Aptitude Verbal test; 2. College Entrance English Essay test; and 3. Yale Verbal Reasoning test.

III. Numerical Ability: 1. College Entrance Scholastic Aptitude Mathematical test; and 2. Yale Quantitative Reasoning test.

IV. Spatial Ability: 1. Yale Spatial Visualization test; and 2. Yale Mechanical Ingenuity test.

Using only Yale undergraduates as subjects restricts the range of

ability sampled.[1] Probably no member of the present group has a
verbal IQ as low as 115.   The range of ability sampled is less restricted
than at first might be supposed, however; the high levels are very well
represented.   Moreover, some of the tests which are not relevant to
general academic achievement, e.g., the measures of spatial ability, may
include a very wide range of scores.   In general it may be claimed that
using a variety of tests which sample relatively homogeneous, specifiable
abilities results in less range restriction than would result in using one
general ability score, e.g., an IQ.

There are two sets of considerations to be observed in generalizing from
the results of the present study: a. If no significant relationships are
found in the present sample, it is unlikely that important linear relation-
ships would be found in a more heterogeneous sample; and b. If the
relationships in the present sample are highly significant, it is possible
that they would have a practical predictive value in a more heterogeneous
sample.

An answer to the two questions raised in the introduction calls for an
examination of the relationships between each of the nine mental ability
measures and each of eighteen Rorschach categories.[2]

Since there were nine mental tests to be correlated with eighteen
Rorschach categories (a total of 162 determinations), it was decided first
to make the simplest preliminary examination of each possible relation-
ship, and subsequently to make a thorough study of the promising rela-
tionships.   For this purpose the 10 highest and 10 lowest people in each
mental test distribution were selected.   This provided nine different
sets of high and low standing students.   Scores on each Rorschach
scoring category were obtained for the high and low standing groups
for each test.

Analysis of Data

Table 1 shows the average number of Rorschach responses for the ten
students who scored highest and for the ten students who scored lowest
on each of the tests.   The two measures of scholastic ability are not
tests; one is merely a grade average and the other is a prediction based

---

[1] This restriction does not preclude the possibility that these tests can show high
correlation in a sample of Yale undergraduates.   Some of the above tests have inter-
correlations as high as .70.

[2] These are: 1. W, Whole Blot; 2. D, Large Usual Detail; 3. d, Small Usual Detail;
4. Dd, Unusual Detail; 5. S, White Space; 6. M, Human Movement; 7. FM, Animals
in Action; 8. m, Abstract or Inanimate Movement; 9. k, Shading on a Three Dimen-
sional Expanse projected on a two dimensional plane; 10. K, Shading or Diffusion;
11. FK, Shading in Three Dimensional Expanse in Vista or perspectus; 12. F, Form
only; 13. Fc, Shading or Surface Texture; 14. c, Shading and Texture; 15. C', Achro-
matic Surface Color; 16. FC, Definite Form with Color; 17. CF, Color with Indefinite
Form; and 18. C, Color only.

on a weighted combination of secondary school grades and mental test scores. With the exception of first semester average, all of the mental measurements show evidence of a positive relationship with the number of Rorschach responses. None of these differences was statistically significant when a t test was made.

The total number of Rorschach responses was found, upon inspection, to be positively skewed for all groups, thus showing a t test of the differences between the means of the total number of responses to be inappropriate. As a consequence the logarithm of each subject's total number of responses was found. The distributions of the logarithms were roughly symmetrical in form, and the t test was repeated based on the logarithms of the total number of responses.

Table 1

The Average Number of Rorschach Responses for the 10 Students Scoring Highest and the 10 Scoring Lowest on Each of the Nine Mental Ability Measurements

Test		Average Number of Rorschach Responses		
		10 Highest on Test	10 Lowest on Test	Difference
I.	Scholastic			
	1. First Semester Average	32.2	35.0	−2.8
	2. General Scholastic Prediction	51.1	35.2	15.9
II.	Verbal			
	1. Scholastic Aptitude Verbal	38.4	33.8	4.6
	2. English Essay	49.4	40.6	8.8
	3. Verbal Reasoning	46.4	38.4	8.0
III.	Numerical			
	1. Scholastic Aptitude Mathematical	48.9	38.1	10.8*
	2. Quantitative Reasoning	42.6	38.5	4.1
IV.	Spatial Ability			
	1. Spatial Visualization	39.2	36.6	2.6
	2. Mechanical Ingenuity	41.1	31.1	10.0

* Difference between logarithms significant at the 5% level.

Only one test, Mathematical Aptitude, showed a difference significant at the five per cent level. The fact that all of the differences (with one exception) are positive, indicates a probably positive relationship between some of the mental tests and the total number of Rorschach responses. The trends (with the possible exception of the one significant difference), are too slight to afford any qualitative evaluation of the pattern of ability and Rorschach response relationships. Considering the great difference

between the low and high level ability groups, the findings offer little support for the practice of using the total number of Rorschach responses as an evidence for mental ability among individual college students.

Because of the indication of a slight positive relationship between total number of responses and measures of mental ability, the location and determinant scores for each individual were expressed as a per cent of his total number of responses.[3] Both the raw scores and the per cent of total scores were analyzed. Despite the large number of differences examined, very few trends were discovered and almost none of them was significant. Only the promising trends will be presented in the following paragraphs.

Table 2

A Comparison Between Pairs of High and Low Scoring Groups on the Basis of Both the Number and the Per Cent of Human Movement Responses

Test		High Group	Low Group	Difference	
		No. M	No. M	No. M	% M
I.	Scholastic				
	1. First Semester Average	5.3	3.6	1.7	4.5
	2. General Scholastic Prediction	7.9	4.6	3.3	1.9
II.	Verbal				
	1. Scholastic Aptitude Verbal	8.8	4.2	4.6	.4
	2. English Essay	8.2	8.1	.1	2.6
	3. Verbal Reasoning	8.1	4.6	3.5	3.7
III.	Numerical				
	1. Scholastic Aptitude Mathematical	8.3	4.3	4.0	1.8
	2. Quantitative Reasoning	8.2	3.5	4.7	.7
IV.	Spatial Ability				
	1. Spatial Visualization	6.5	4.3	2.2	4.1
	2. Mechanical Ingenuity	7.9	5.4	2.5	5.0

Table 2 indicates the nature of the relationship between the number of Human Movement (M) responses and the mental ability measures. It is apparent that there is a general tendency for the number of Human Movement responses to be positively related with mental ability measurements. This tendency is not wholly due to the fact that mental ability is slightly related to total number of responses; this is indicated by the consistent positive differences between the groups in *per cent* Human

[3] In his study of the relationships between Bellevue-Wechsler scores and Beck's Rorschach scoring factors, Wishner (8) makes no adjustment for the manner in which some of the scoring factors may be influenced by the total number of responses (R). This is regrettable because his data suggest that the validity he claims for Z could be largely if not entirely due to R.

Table 3

A Comparison Between Pairs of High and Low Scoring Groups on the Basis
of the Per Cent of Whole Responses

Test		High Group %W	Low Group %W	Difference %W
I.	Scholastic			
	1. First Semester Average	42.3	40.2	2.1
	2. General Scholastic Prediction	34.2	42.9	−8.7
II.	Verbal			
	1. Scholastic Aptitude Verbal	34.2	38.4	−4.2
	2. English Essay	33.5	27.2	6.3
	3. Verbal Reasoning	29.0	35.0	−6.0
III.	Numerical			
	1. Scholastic Aptitude Mathematical	29.5	34.9	−5.4
	2. Quantitative Reasoning	32.5	30.5	2.0
IV.	Spatial Ability			
	1. Spatial Visualization	33.0	32.8	.2
	2. Mechanical Ingenuity	24.5	37.0	−12.5

Table 4

Evidence for Relationship Between Tendency to Give Achromatic Color Responses and
Tendency to be in the High or Low Scoring Groups for Each of the Mental Tests

		χ^{2}*	P
I.	Scholastic		
	1. First Semester Average	.95	.30
	2. General Scholastic Prediction	5.47	.02
II.	Verbal		
	1. Scholastic Aptitude Verbal	7.2	.01
	2. English Essay	.8	.30
	3. Verbal Reasoning	5.05	.02
III.	Numerical		
	1. Scholastic Aptitude Mathematical	5.02	.02
	2. Quantitative Reasoning	.22	.70
IV.	Spatial Ability		
	1. Spatial Visualization	.833	.30
	2. Mechanical Ingenuity	1.97	.20

* Without Yates correction.

Movement responses. The number of Human Movement responses like the total number of responses proved to be positively skewed; as a consequence a t test was based on the logarithms of the per cent of Human Movement scores. None of the differences proved to be significant at the five per cent level.

The number of whole responses showed little promise of being related with the mental ability scores. Table 3 shows the ambiguous finding for per cent whole responses.

Only one of the other scoring categories for determinant or location factors showed evidence of being related with mental ability. This was the number of achromatic color responses (C').

Since for any individual the number of achromatic color responses (C') was small, no t test was feasible and no correction for the influence of the total number of responses on the number of achromatic color responses was made. The reliability of the relationship between the number of achromatic color responses and mental ability scores was examined by means of a χ^2 test of independence, table 4.

Discussion

The experimental findings are discussed with respect to the two questions to which the experiment is specifically relevant: 1. What is the order of any linear relationship between a Rorschach response category and test evidence for mental ability? 2. What is the pattern of linear relationships between various classes of mental ability and various Rorschach response categories?

With respect to the first question, it is apparent that no linear relationship of sufficient strength to justify *individual* evaluation exists between any type of mental ability sampled and any one of the usual Rorschach location or determinant scoring categories. The qualification "linear" is offered because it is possible that at a low level of ability a more appreciable relationship exists between mental ability and frequency of responses in certain of the Rorschach categories. Such discontinuous or curvilinear relationships have not proved to be important in mental ability studies, however.

Because of the paucity of evidence for reliable relationships between mental ability and the selected Rorschach response categories, the second question becomes irrelevant. The slight trends observed give no hint that certain types of responses are correlated with certain types of ability.

Obviously the present findings do not preclude the possibility that the Rorschach may be used in some manner or other to predict some aspect of mental ability. The present study does indicate the limited value of Rorschach location and determinant categories as evidence for

mental ability. This suggests that the accuracy of Rorschach perceptions (form level ratings (4)) and other cues are the primary basis for any valid appraisal of mental ability. Such cues are not particularly objective; their evaluation is informal and not well standardized. The reliability of form level ratings accrues from the consistency of the examiner,[4] and their validity is dependent upon his judgment. Thus it appears that the most formal and objective aspects of a Rorschach protocol (the location and determinant category scores) have almost no validity. The remaining factors (form level and the other purely qualitative cues) are likely to be unreliable or, at best, to possess a reliability which is more a characteristic of the examiner than of the Rorschach procedure. Concerning the possible validity of accuracy of perceptions as an evidence for mental ability, it is of interest that Beck's (1) F plus % (probably more reliable than Klopfer's form level ratings) was found by Hertz (2) to be correlated with mental ability; Wishner (8) could not confirm this, however.

Summary and Conclusions

The present study is an examination of the relationships between measures of scholastic, verbal, numerical, and spatial abilities and the commonly used Rorschach scoring categories for location and determinant factors. The subjects were a sample of sixty-eight Yale students. The findings may be summarized in the following manner:

1. Although the total number of responses, the number of whole responses, or the number of Human Movement responses is often used as a part of the procedure for estimating mental ability from Rorschach protocols, in the present sample none of them has sufficient validity to to justify use for distinguishing between individual college students of different levels of ability.

2. If the relationships between any Rorschach location or determinant category and any of the types of mental ability used in the present study is linear, the evidence from this sample indicates that their value for predicting individual mental ability is so scant as to make their use at any ability level uneconomical and misleading.

3. The present negative or negligible findings do not preclude the possibility that some examiners, employing other aspects of the protocol or clues not gained from the Rorschach responses, may arrive at valid estimates of some sort of mental ability.

4. There is evidence of a slight tendency for the total number of Rorschach responses (R) to be positively correlated with several measures of mental ability. This finding requires that all of the other comparisons

[4] This was recognized by Wishner (8).

had to be corrected for differences in total number of responses in order to eliminate the spurious effect of a third variable.

5. Two of the Rorschach scoring categories based on the determinants of a response (the color, shading, or movement factors) show evidence for a slight positive relationship with measures of mental ability. They are the number of Human Movement responses and the number of achromatic color responses.

6. None of the Rorschach categories based on the location factor (portion of the card used in forming a response), is related with any of the measures for mental ability. Significant trends were absent not only among the skewed raw scores but among their logarithms as well.

References

1. Beck, S. J. *Rorschach's test: Vols. I and II.* New York: Grune and Stratton, 1945.
2. Hertz, M. R. The Rorschach Ink Blot Test: A historical summary. *Psychol. Bull.,* 1935, 32, 33–66.
3. Hertz, M. R. Rorschach norms for an adolescent group, *Child Develop.,* 1935, 6, 69–76.
4. Klopfer, B., and Davidson, H. H. Form level rating. *Rorschach Res. Exch.,* 1944, 8, 164–177.
5. Klopfer, B., and Kelley, D. M. *The Rorschach Technique.* New York: World Book, 1942.
6. Rapaport, D. *Diagnostic psychological testing: Vols. I and II.* Chicago: Year Book Pub., 1945.
7. Rorschach, H. *Psychodiagnostics.* Berne: Hans Huber, 1942.
8. Wishner, Julius. Rorschach intellectual indicators in neurotics. *Amer. J. Orthopsychiatry,* 1948, 18, 265–279.

Modification of Academic Performance
through Personal Interview *

Alex C. Sherriffs

University of California

Among the many problems facing university teachers today is that of the large class. Each year finds a greater proportion of university courses with enrollments in the hundreds. Some of the larger universities report over a thousand students in the beginning courses of certain popular fields.

Most instructors feel that the large class is an educational hazard. The negative aspects perhaps most frequently cited include the minimal opportunity for student participation during lecture hours, the necessity for using recognition type examinations which usually do not call for the integration of course material, serving only the purpose of providing a basis for grading students, and the essential lack of contact between individual students and the course instructor.

It is with one phase of this latter aspect that this paper is concerned. The experiment reported here is intended to throw some light on the significance of the contact of individual students with their instructor, especially in the situation of the large class.

This experiment was formulated on the basis of three hypotheses. These were: (1) that those students of a large class who felt themselves to be known as individuals to their instructor would demonstrate more effective learning of course material than would their fellows not so known; (2) that there would be demonstrable individual differences in the effects of being known to the instructor; and, (3) that such individual differences could be predicted with some accuracy.

Procedure

To test the first hypothesis, it was decided to subject a random sample of students in a large class to a sixty-minute interview by their instructor during the week following the first midterm examination. Scores on this examination would serve as a baseline against which to compare the performance of these students on examinations following the interview. The remainder of the class would serve as the control group.

To test the second and third hypotheses, judgments would be obtained on the students interviewed as to certain personality variables. These variables would be characteristics considered likely to modify the effect of an interview contact by the instructor. Students high on such characteristics would be compared in their performance on examinations taken after the time of the interviews with those low on these characteristics, and both of these groups would be compared with the non-interviewed students of the class.

The Subjects

The class chosen for this experiment was the beginning survey course in psychology at the University of California. This class was chosen simply because of its availability and its large enrollment. The experimenter was the instructor, and some 257 students were registered and took the examinations throughout the course.

This course is open only to those not intending to major in psychology. The students were all freshmen and sophomores who ranged in age from 17 to 24, with a mean age of 19.0 for the group.

The course extended over a sixteen-week period, with three lectures each week, and with objective recognition type examinations. Midterms were administered during the fifth and tenth weeks, and the final examination was held at the end of the sixteenth week. All students in the course were required to serve as subjects for two hours of laboratory experiment during the semester. The interview for the subjects of the present investigation counted as one of the regular laboratory hours.

A sample of thirty-four students was selected for the experimental group by including every eighth student on the class roll. Check indicated the sample to be representative of the class as a whole on the variables of age, sex, and academic major. Of importance was the fact that the distribution of scores made on the first midterm by this group was highly similar to that made by the rest of the class (See Table 1).

Table 1
Comparison of the Experimental Group with the Remainder of the Class on the First Midterm Examination

	Mean	S.D.	*t*
Experimental Group (N = 34)	49.5	4.82	
			.17
Remainder of Class (N = 223)	49.3	6.80	

The Interview and the Personality Variables. Since the main function of the interview was to cause each student of the experimental group to feel that he was known to the instructor as a definite individual, and since it was also desired to gain information concerning each subject so as to make certain personality judgments, the interviews were directed at procuring life history and attitudinal material. The instructor carefully avoided discussion of material of the course or of the student's reaction to the class. The explanation given the student for the interview was that it was desired to know as much as possible of the interests and backgrounds of those enrolling in this course.

The personality variables chosen from among those likely to have significance in relation to the effect of the interview on the student's academic performance follow:

1. *Self-tension.* The amount of tension felt by the student as to his own adequacy and worth.
2. *Family-tension.* The amount of tension felt by the student in his family relations, in regard both to parents and siblings.
3. *Social-tension.* The amount of tension felt by the student in his social relations.
4. *Over-all tension.* The general level of tension and anxiety under which the student functions, taking into account the above three areas.
5. *Achievement need.* The importance to the student of high academic grades.
6. *Affection need.* The importance to the student of receiving a constant supply of warmth and affection from others.
7. *Praise need.* The importance to the student of praise and recognition from others.

Obviously these variables would not be completely independent, but it seemed that their individual meaning was sufficiently separate to be useful for this study. Intercorrelation of measures of variables was no handicap so long as a true relationship was represented. The real difficulty lay in not having independent observers to obtain the different measures. By the nature of the study only the course instructor could interview the group of subjects. The amount of intercorrelation resulting from "halo effect" operating on the one interviewer cannot be determined.

Five-point rating scales with defined points were utilized for the judgments of the four tension variables. These were rating scales previously found to be useful by the writer.[1] Seven-point rating scales,

[1] Sherriffs, A. C. The "Intuition Questionnaire": A new projective test. *J. abnorm. soc. Psychol.*, 1948, **43**, 326–337.

with the points defined in terms of probability of occurrence, were employed for the ratings of the subjects on the three "need" variables. These rating scales follow closely the method outlined by Murray.[2]

Results [3]

1. *First Midterm Examination.* The first task was to assess differences on the first midterm examination between the randomly selected sample and the remainder of the class. This examination, it will be remembered, was administered before the members of the experimental group were interviewed. The pertinent data for this comparison appear in Table 1.

This comparison does not suggest that the experimental group was different from the remainder of the class in terms of performance on the first midterm examination.

Table 2

Comparison of the Experimental Group with the Remainder of the Class in Terms of the Shifts in Scores from the first Midterm to Subsequent Examinations

	Midterm I to Midterm II			Midterm I to Final		
	Mean	S.D.	t	Mean	S.D.	t
Experimental Group (N = 34)	+6.6	5.52		+80.4	9.99	
			2.39*			1.88
Remainder of Class (N = 223)	+4.3	5.31		+76.3	12.08	

* Significant at the 2 per cent confidence level.

2. *Performance on Second Midterm Examination and on Final Examination as Compared with Performance before the Interview.* Suggestions as to the effect of the interview on the performance of the experimental group of subjects come from comparisons of this group with the remainder of the class in their functioning on later examinations relative to their functioning on the first, pre-interview, midterm. The mean

[1] Murray, H. A. *Explorations in personality.* New York: Oxford University Press, 1938.

[3] All estimates appearing in this paper of the significance of the differences between means are based on the t test. Comparisons involving a two part split of the thirty-four subject experimental group require a t of 2.04 to be significant at the 5 per cent confidence level, and a t of 2.74 to be significant at the 1 per cent confidence level. Those comparisons which involve the 223 subjects not included in the experimental group require a t of 1.96 to be significant at the 5 per cent level, and a t of 2.58 to be significant at the 1 per cent level.

differences in points scored on the first midterm examination and those scored on the second midterm and those scored on the final examination are presented in Table 2. The variabilities of the shifts in performance, and the significance of the differences between the shifts of the experimental group and of the remainder of the class are also shown.

These comparisons suggest that the interviewed group of students improved more than did the remainder of the class in their performance on the second midterm examination held four weeks after their contact with the instructor. The difference in improvement is significant at the 2 per cent confidence level. The difference still favors the experimental group at the time of the final examination some ten weeks later, but this latter difference is not significant at the 5 per cent level.

Relationship of Rated Personality Variables to Effects of Interview on Performance

Distributions of ratings were made for each of the seven personality variables to be studied. These distributions were then considered separately and split in each case so as most closely to accomplish a 50–50 division of the subjects on that particular variable. Comparisons could then be made of the examination performance of those subjects rated higher on each variable with the examination performance of those rated lower.

1. *Relationship of Rated Personality Variables to Performance on the First Midterm Examination.* The means and standard deviations of the scores made on the *first* midterm examination by those rated higher and those lower for each personality variable are shown in Table 3. Estimates as to the significance of the differences between these means are also indicated.

Table 3

Performance on the first Midterm Examination of those of the Experimental Group Rated Higher and of those Rated Lower on Seven Personality Variables

Variable	Higher Ratings			t	Lower Ratings		
	N	Mean	S.D.		N	Mean	S.D.
Self Tension	16	48.1	5.82	1.62	18	50.8	3.24
Family Tension	14	48.3	5.54	1.25	20	50.4	4.03
Social Tension	17	48.1	5.06	1.73	17	50.9	4.09
Overall Tension	13	47.5	4.77	2.03*	21	50.8	4.39
Achievement Need	13	50.0	4.77	.44	21	49.2	4.83
Affection Need	10	45.3	4.67	3.22**	24	51.3	4.38
Praise Need	12	46.8	4.88	2.65*	22	51.0	4.05

* Significant at the 5 per cent confidence level.
** Significant at the 1 per cent confidence level.

These comparisons reveal that the students rated as most tense, generally, and in each of the three areas of tension, did less well on the first midterm than did their fellow students who were rated as less tense. Those students judged most strongly to need affection and praise did less well than did those judged to have less of these needs. In the case of the achievement need we find that those students with higher ratings performed better on the examination. The differences between means are significant at the 5 per cent level of confidence in the cases of overall tension and need for praise, and at the 1 per cent level in the case of need for affection.

Table 4

Relationship to Rated Personality Variables of Shifts in Scores
from First Midterm to Subsequent Examinations*

| | Midterm I to Midterm II** | | | | Midterm I to Final** | | | |
| | Higher Ratings | | Lower Ratings | | Higher Ratings | | Lower Ratings | |
Variable	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Self Tension	8.8	4.78	4.7	5.45	83.5	9.12	77.7	9.92
Family Tension	8.9	4.79	5.0	5.43	80.9	10.30	80.1	9.72
Social Tension	8.3	4.93	4.9	5.59	82.0	9.32	78.8	10.36
Overall Tension	8.7	4.37	5.3	5.77	81.1	10.01	80.0	9.94
Achievement Need	6.6	3.81	6.7	6.35	83.8	10.56	78.3	8.97
Affection Need	10.8	4.12	4.9	5.08	83.2	8.40	79.3	10.35
Praise Need	8.8	4.76	5.5	5.56	84.3	8.76	78.3	9.99

* The N's for the subgroups of subjects represented in this table may be found in Table 3.

** All shifts are positive for in all cases means are higher on the second midterm and on the final examination than on the first midterm.

The implications of these findings would seem to be that degree of tension and amount of need for affection and for praise are related to examination performance by students in large classes. One might hazard guesses as to the further meaning of these data, for example, the relation of these tensions and these needs to academic performance generally, regardless of class size, and the deeper meaning of the presence of high and low tension and strong and weak needs in regard to personality structure and function.

2. *Relationship of Rated Personality Variables to Performance Occurring after Contact between Student and Instructor*. It was necessary to find a measure of the effect on performance of contact with the instructor, at the same time relating this effect to the seven personality variables being

investigated. The sole use of scores on the second midterm examination and on the final examination would be inadequate because of the findings presented in the previous section. Such scores would be ambiguous in meaning for our purposes because of the fact that the personality variables were shown to be related directly to performance. This relationship would somehow have to be taken into account before the effect of the instructor contact could be isolated.

The measure best serving the purposes of this study was that of the *differences* in performance before and after contact with the instructor as related to ratings on the personality variables. Data on such shifts in mean performance from the first midterm examination to the second midterm examination and to the final examination will therefore be presented.

Table 5

Significance of the Difference in Scores on Examinations Taken Before and After
Interview Contact with Course Instructor

	Midterm I to Midterm II			Midterm I to Final		
	High on Variable vs. Low	High on Variable vs. Class	Low on Variable vs. Class	High on Variable vs. Low	High on Variable vs. Class	Low on Variable vs. Class
Variable	t	t	t	t	t	t
Self Tension	+2.21*	+3.27**	+ .35	+1.72	+2.33*	+ .47
Family Tension	+2.11*	+3.20**	+ .59	+ .24	+1.39	+1.34
Social Tension	+1.80	+3.01**	+ .50	+ .91	+1.90	+ .83
Overall Tension	+1.75	+2.93**	+ .87	+ .30	+1.39	+1.35
Achievement Need	− .02	+1.57	+1.94	+1.50	+2.19*	+ .73
Affection Need	+3.16**	+3.82**	+ .54	+1.04	+1.78	+1.15
Praise Need	+1.68	+2.85**	+1.00	+1.67	+2.24*	+ .75

  * Significant at the 5 per cent confidence level.
  ** Significant at the 1 per cent confidence level.

In Table 4 the means and standard deviations of the differences in scores on examinations taken before and after interview contact with the instructor are presented. The experimental group is broken down into those students rated higher and those rated lower on each personality variable.

The significance of the differences in shifts in examination scores between: (1) those students high on each variable and those low; (2) those high on each variable and the remainder of the class; and (3) those low on each variable and the remainder of the class, were then calculated. Table 5 summarizes the resulting t's.

The data summarized in Tables 4 and 5 suggest that:

1. The effect of a single interview contact by the individual students of the experimental group with their course instructor was not uniform. There was significantly (at the 1 per cent level) more effect on those students rated higher on self tension, family tension, social tension, over-all tension, affection need, and praise need than on those rated lower—when one compares the performance of these students with that of the non-interviewed students of the class.

2. The effect of the interview contact diminishes over the ten-week period before the final examination, holding up (at the 5 per cent level) for only three out of the seven variables, and only then in the case of those subjects judged relatively high on these variables. Nonetheless, comparison of the scores of the subgroups of the interviewed subjects shows them all to have higher mean scores than the mean score attained by the remainder of the class as a whole.

The limitations of this study in terms of numbers of subjects in the experimental subgroups, lack of controls over the possibility of the operation of "halo effect" on the personality ratings, and the fact that the data obtained are from one class at one university and in relationship to one instructor do not allow for definite generalizations to students, classes, and instructors the world over. However, the writer feels the results of this study to be evidence for the value of personal interviews with students in large classes. These results further suggest that some students are handicapped in their performance by the lack of student-teacher contact and the lack of individualization felt when an "unknown" member of a class. This study points to the possibility of discovering those students who need most and who would profit most from individual attention. Conversely, it indicates the possibility of screening those students who would be handicapped but little by membership in a large class insofar as lack of contact with the instructor is concerned. It is of particular interest to the writer that the significant improvement in examination performance made by students following the interview with their instructor was made after a single contact, and a contact of only one hour. The results of a study on the effects of continued conferences might truly be exciting.

Vocabulary Item Difficulty and Word Frequency

James J. Kirkpatrick and Edward E. Cureton

University of Tennessee

In constructing a vocabulary test, it is desirable in many cases to arrange the items of the experimental edition in approximate order of difficulty. Test constructors often try to do this by arranging them in the order of frequency of occurrence of the key words. Questions immediately arise concerning the validity of this procedure, and the possibility of improving it by the use of direct judgments. A study designed to throw some light on these matters was made, using the 100 four-choice vocabulary items of the Army General Classification Test, Forms 1a and 1b. The difficulties of these items, as reported by the Staff, Personnel Research Section, The Adjutant General's Office (4), are given in terms of the percentages of correct responses made by the soldiers in the experimental tryout samples. The Form 1a sample included 400 cases; the Form 1b sample, 218. The difficulty values of the Form 1b items were adjusted to make them comparable to those of Form 1a. The frequency of the key word (the stem-word of the item or the correct answer, whichever was least frequent[1]) was taken as the frequency value. The frequencies were taken from *The Teachers Word Book of 30,000 Words* (6). This word book reports the frequencies of words in terms of the number of occurrences per million running words, for the 19,440 words which are encountered at least once per million; and in number of occurrences per eighteen million running words, for those which are encountered less frequently than once per million, but more frequently than once per four million. The 952 words which appear 50 to 99 times per million are lumped together and simply labeled A; the 1,069 words which appear 100 or more times per million are all labeled AA.

The number of different words at each frequency-of-occurrence level forms a J-shaped distribution, which can be fairly well represented by an exponential function. In order to obtain a more or less symmetrical distribution of frequency measures, the common logarithms of the frequencies of occurrence were grouped into equal intervals. The interval-width was determined by the fact that all words in the A-group in the word book (50–99 per million) had to go into one group. For the less

[1] In six of the 50 items of Form 1a, and in one of the 50 items of Form 1b, the correct answer was of lower frequency of occurrence than the stem-word, according to the word book.

frequent words, the numbers of occurrences per eighteen million were divided by eighteen, and the quotients were rounded off to one decimal. This procedure gave eleven groups containing the following frequencies:

Group	Range of Frequencies per Million	Number of Words in Group	Number of Words in ACCT 1a and 1b
1	100+ (AA)	1069	2
2	50–99 (A)	952	5
3	26–49	1256	3
4	13–25	1865	18
5	7–12	2506	10
6	4–6	2638	21
7	2–3	3945	15
8	.8–1	**	13
9	.4–.7	**	10
10	.2–.3	**	2
11	not in list	**	1

** Not reported in word book.

We were also interested in the possibility of *improving* on these frequency-estimates of difficulty by the use of direct judgment (not, in this case, in testing the validity of direct judgment *per se*). Each of the 100 items was therefore typed on a 3″ by 5″ card, and the frequency group recorded on the face of the card. Each judge was presented with the eleven groups of cards, informed concerning the basis of the grouping, and requested to rearrange the cards among the eleven groups so that they would be more nearly in the order of their "true difficulty" (defined as the probability that the average American soldier in World War II would get the right answer). They were required to keep to exactly eleven groups, but were *not* required to keep the same number of cards in each group as the number given by the frequency-count. Judgments were secured from five English instructors, each of whom worked independently.[2] The sum of the five group-allocations was computed for each item. These sums ranged from the minimum possible, 5, to the maximum possible, 55, the larger numbers representing greater judged difficulties.

A third estimate of difficulty consisted simply of a count of the number of syllables in the key word of each item (see Flesch, 2). These numbers ranged from one to five.

The validities of the three methods of estimating item difficulties were determined by correlating these estimates with the criterion given

in the Army study.   The correlations are as follows:

Criterion with frequency	.47
Criterion with judgment	.71
Criterion with syllable-count*	.20
Frequency with judgment	.81

* Sheppard's correction applied to standard deviation of syllable-count.

The last correlation reported above is not a validity coefficient, and it is spuriously high because the judges knew the frequency groups to begin with.   It was computed because it was needed in testing the significance of the difference between the first two criterion correlations.

Inspection of these correlations immediately suggests the marked superiority of the frequency-plus-judgment technique, and the equally marked inferiority of the syllable-count technique.   It seems reasonable to suggest, on the basis of this latter finding, that the authors of "read-biblity" formulas investigate the relative merits of counting syllables as against having a single judge estimate word-difficulties.   Since five judges participated in this study, and since they started with knowledge of the frequency-counts, the outcome of such studies cannot, of course, be predicted.

The significance of the difference between the correlations of frequency and judgment with the criterion was evaluated by Hotelling's adaptation of Student's t-test (3).   The value of t was 5.5, which is clearly significant at the .001 level.

Applying the Fisher z-transformation to the correlation of .71 between difficulty and frequency-plus-judgment, it was found that the chances are 19 to 1 that its "true" value lies between .60 and .80.

A second study, concerned with difficulty and frequency only, was based on a set of items consisting of word-pairs to be marked S, O, or N, depending on whether their meanings were the same, the opposite, or neither.   Three hundred such items were administered to about 500 high school seniors.   On the basis of total scores on the 300 items, the top 100 and the bottom 100 examinees were selected as criterion groups, and 68 items were discarded on the basis of failure to discriminate between these two groups.   The difficulty of each of the remaining 232 items was taken as the per cent correct in the combined group of 200.   The frequency value was taken as the ordinal thousand, in *The Teachers Word Book of 20,000 Words* (5), of the least frequently occurring word in the pair.   For this set of 232 items, the correlation between difficulty and frequency was .56.

Davis (1) has reported low correlations between item difficulty and stem-word frequency, as given by *The Teachers Word Book of 20,000*

Words (5), for three types of vocabulary items from the Cooperative English Test. Two of these item types require the examinee to supply a word which matches a given definition. The factor-analysis literature (7, e.g.) suggests that such items measure "verbal fluency" to some considerable degree, whereas items of the types reported in our own studies measure mainly "verbal relations." The third item type studied by Davis was apparently more like those of the Army General Classification Test: a stem word followed by five alternatives from among which the examinee was required to select the synonym of the stem word. Using 208 items of this type, Davis found a correlation between difficulty and frequency of only .10.

It is quite possible that the superficial similarity between the Cooperative Vocabulary Test items and those of the Army General Classification Test is considerably greater than their actual content similarity. The Cooperative items were designed to measure *precision* of knowledge of fairly common words. Davis criticizes the practice of including rare words to provide difficult items in vocabulary tests. He says (1, pp. 71–2), "The difficulty of a multiple-choice vocabulary item for a given group of subjects is dependent on two main factors: first, the per cent of the group that could define the word correctly if asked to state its meaning; and, second, the degree of discrimination required to distinguish between the correct answer and the incorrect answers, or decoys, in the item. The importance of this second point has often been overlooked with unfortunate results. Test constructors have built items to test for knowledge of words like 'syzygy' or 'umbel.' Such words have virtually no practical value except to specialists in certain learned professions; hence, they reduce the real validity of general vocabulary tests, but they have been included to provide very difficult items in vocabulary tests that are not made up of items in which the decoys have been chosen with care and ingenuity so that they differ only slightly, though incontestably, from the correct meanings of the words being tested."

The force of this argument would appear to depend on the purpose for which the test is designed. We can see no objection to designing vocabulary tests to measure *range* of word knowledge at a low level of discrimination, as well as *precision* of knowledge of fairly common words. The same-opposite-neither test is clearly of the former type. The Cooperative Vocabulary Test is of the latter type. The vocabulary items of the Army General Classification Test fall somewhere between these two extremes. Examination of its item-alternatives suggests that it is probably more nearly a range test than a precision test.

Comparing the three correlations between frequency and difficulty, there appears to be a fairly definite trend. For the precision-type Co-

operative Vocabulary Test the correlation is .10. For the vocabulary items of the Army General Classification Test, it is .47. For the same-opposite-neither test, it is .56. It seems reasonable to suggest, as a hypothesis if not as a conclusion, that the nearer a vocabulary test comes to being a measure of *range* rather than *precision* of word knowledge, the higher will be the correlation between the frequency values of its key words and the difficulties of its items. Moreover, the estimates of difficulty based on frequency can be improved markedly by the use of direct judgment.

References

1. Davis, F. B. The interpretation of frequency ratings obtained from "The Teachers Word Book." *J. educ. Psychol.*, 1944, *35*, 169–174.
2. Flesch, R. A new readability yardstick. *J. applied Psychol.*, 1948, *32*, 221–233.
3. Hotelling, H. The selection of variates for use in prediction with some comments on the problem of nuisance parameters. *Annals of Math. Statist.*, 1940, 11, 3, 271–283.
4. Staff, Personnel Research Section, The Adjutant General's Office, The Army General Classification Test, with special reference to the construction and standardization of Forms 1a and 1b. *J. educ. Psychol.*, 1947, 38, 385–420.
5. Thorndike, E. L. *The teacher's word book of 20,000 words.* Bureau of Publications, Teachers College, Columbia University, 1931.
6. Thorndike, E. L., and Lorge, I. *The teacher's word book of 30,000 words.* Bureau of Publications, Teachers College, Columbia University, 1944.
7. Thurstone, L. L. *Primary and mental abilities.* Psychometric Monograph No. 1, University of Chicago Press, 1938.

Influence of Prestige Suggestion on the Answers of a Personality Inventory *

Joseph F. Donceel, Benjamin S. Alimena and Catherine M. Birch

Fordham University

The following investigation was inspired by an experiment performed in 1933 by two German psychologists, H. Krüger and K. Zietz (1). They composed an artificial personality description and told each of 30 subjects that this description was based on a graphological analysis of their handwriting and on a study of their horoscope. All the subjects accepted this one standard description as a good analysis of their personality; some were amazed at its accuracy; not a single subject rejected the diagnosis as a whole.

Among the possible explanations of this surprising result, the authors noted: the fact that the subjects do not know their own personality structure; their suggestibility; the vague and ambiguous character of many of the statements used in the personality description.

The purpose of the present experiment was to find out to what extent subjects would accept as applying to them a personality description obtained by mere chance, even when the statements used in this description were not vague and ambiguous, and even when no effort had been made to avoid the contradictions which derive from a random accumulation of statements.

First Experiment, Using Mild Suggestion. The subjects for the first experiment were 34 students in a psychology class for adults, both men and women, ranging in age from 20 to 55 and in education from four years completed in High School to two years completed in College. The subjects were asked to hand in a specimen of their hand-writing, and they were told that the experimenter would have it analyzed by a graphologist, and would give them a detailed description of their personality, based on this analysis.

In fact, the experimenter just took for each subject a Bernreuter Personality Inventory and answered its 125 questions at random. The questions were matched with the 125 first figures of a table of random numbers; when the figure for a certain question was even, that question received a "yes" answer; when the figure was odd, that question received a

"no" answer. A week after the handwriting samples had been received, the Bernreuter Inventories were given to the subjects with the affirmative or negative answers, and the subjects were asked to check each of the statements, and to indicate whether they agreed or disagreed with the answer.

From chance alone we expect a number of agreements averaging 50 per cent, that is, an average agreement with 62.5 of the 125 suggested answers. Any number of agreements higher than 73 would occur by chance alone only 5 times out of 100, whereas a number of agreements higher than 77 would occur by chance alone only once in 100 times.

The number of agreements of the 34 subjects ranged from 60 to 100. The average number of agreements was 78 with a standard deviation of 9.41. The results of 4 subjects excluded the null hypothesis at the 5 per cent level of confidence, whereas the results of 15 more excluded the null hypothesis at the 1 per cent level of confidence. In other words, 19 out of our 34 subjects agreed with the suggested statements more often than could be explained by chance alone. They gave evident signs of suggestibility in their self-analysis.

Second Experiment, Using Stronger Suggestion. The second experiment employed stronger suggestion. Fifty subjects were used, 25 men and 25 women, ranging in age from 18 to 33 and in education from two years completed in High School to completion of Graduate Studies. Here again a Bernreuter Personality Inventory was answered for each of the subjects by mere chance, just by rolling dice. Each of the subjects was given individually a Rorschach Inkblot Test and a Murray Thematic Apperception Test. Next, allegedly on the basis of these tests, the experimenter answered orally, in the presence of the subject, the 125 questions of a Bernreuter Inventory (that is, gave for each question the answer determined by the dice) and asked the subject to tell whether or not he or she agreed with that answer.

From chance alone we expect an average number of 62.5 agreements. The actual number of agreements ranged from 83 to 125; the average was 111.6 with a standard deviation of 9.16. Since chance alone is excluded at the 1 per cent level of confidence for any number of agreements higher than 77, the null hypothesis was excluded for every one of the 50 subjects.

There was no reliable difference between the amounts of agreements shown by the men and by the women. For the men the average was 112.1 and for the women 111.0.

Every question of the Inventory was answered for the 50 subjects. Therefore, from chance alone, it is expected that 25 subjects will agree with the suggested answer to each question. Any number of agreements for a single question higher than 34 is significant at the 1 per cent level.

We obtained an average agreement per question of 44.6 with a standard deviation of 3.44 and a range of 35–50. Hence, for each single question, effective suggestion could be established at the 1 per cent level of confidence.

If we consider each question individually for the men alone, we find two questions for which the number of agreements is only 17 out of 25. Here chance alone cannot be excluded, even at the 5 per cent level of confidence. These questions are: "Do you ever complain to the waiter when you are served inferior or poorly prepared food?" and "Have you been the recognized leader of a group within the five last years?"

The same applies for three questions of the female group: "Do you frequently argue over prices with tradesmen or junkmen?" (For this question suggestion did not work at all, the percentage of agreements was only 52); "Do people ever come to you for advice?" and "Are you systematic in caring for your personal property?" It will be noticed that these five questions are of a clearly factual nature.

Immediately after the test with suggestion, the experimenter explained to the subjects that the answers which had been presented had been obtained by a mere chance procedure and were therefore without value. He gave each subject an unanswered Bernreuter Inventory and asked him or her to answer all the questions personally. This would yield a measure of the endurance of the suggestion.

Instead of finding the expected average of 62.5 agreements with the previously suggested answers, we found an average of 87.4 agreements, with a standard deviation of 10.2 and a range of 67–109. In 40 out of the 50 subjects suggestion could still be established at the 1 per cent level of confidence. Only 6 subjects were able to shake off the suggestion enough to yield results insignificant even at the 5 per cent level of confidence.

Third Experiment, with Suggested Reversal. In our last experiment the subjects were 49 sophomores attending a Liberal Arts College for Women. This time the subjects first answered the questions of a Bernreuter Inventory in the ordinary way, without any suggestion. Then they were given a group Rorschach Test. Four weeks later the experimenter met each subject individually and told her that, for a certain number of answers, the results of the Rorschach Test contradicted the answers given by the subject. She was asked whether she did not feel that the answer suggested by the Rorschach Test corresponded better with reality. In other words, in this experiment we did more than to suggest a certain answer to the subject; we tried by means of suggestion to make her repudiate or reverse a previously given answer and accept the opposite answer as the true one.

We did not try, of course, to make the subjects change all the pre-

viously given answers; they would have suspected some trick.   Reversal was attempted under suggestion for one third approximately of the answers, for 42 answers taken at random.   Of these 42 suggestions, an average of 26 was accepted, or approximately 60 per cent.   There were, of course, considerable individual differences; the lowest number of accepted reversals was 10 per cent, the highest number was 94 per cent.

These results are highly significant.   It is true that Lentz (2) found that, when subjects were retested with the Bernreuter after a lapse of from one to four weeks, they changed approximately 20 per cent of their original answers.   If we take that amount of change as a measure of the modifications which may be due to the mere lapse of time, we find that the 60 per cent of change discovered in our experiment yields a chi square of 39.67, which is considerably above the 6.64 required for significance at the 1 per cent level of confidence.

Summary

1. The questions of a Bernreuter Personality Inventory were answered for a group of subjects.   These answers, obtained by mere chance, were presented as the results of psychological tests, and the subjects were asked to tell whether they agreed or disagreed with these answers.

2. When mild suggestion was used, 19 out of 34 subjects accepted the answers more often than could be explained by chance alone.

3. When stronger suggestion was used, 50 out of 50 subjects yielded to suggestion.

4. Subjects were also induced, under suggestion, to repudiate 60 per cent of their own answers to the Inventory and to accept as true the opposite answers.

References

1. Krüger, H. and Zietz, K.   Das Verifikationsproblem.   *Z. angew. Psychol.*, 1933, 45, 140–171.
2. Lentz, T. F.   Reliability of opinionaire technique studied intensively by the retest method.   *J. soc. Psychol.*, 1941, 14, 229–256.

A Note on Kahn and Hadley's
"Factors Related to Life Insurance Selling"

S. Rains Wallace, Jr.

Life Insurance Agency Management Association, Hartford, Conn.

In a recent article in this Journal, Kahn and Hadley (1) have reported a study in which it was proposed: "first, to determine the degree of relationship that exists between relative success in the early period of selling life insurance and success at a later period; second, to examine various selling activities with a view to uncovering certain factors which differentiate successful from unsuccessful agents . . . ; third, to investigate further certain personal history items and personality traits already known to correlate with success in selling life insurance, and to analyze other measurable areas of personality, with the aim of increasing the sensitivity of existing selection methods." The authors further assert that "The identification of individuals for whom the likelihood of success is known would not only benefit management, but would, to some extent, minimize feelings of frustration on the part of the agent who, from the outset, may be doomed to failure."

The writer is in full accord with these aims (if dubious of "personality traits *known* to correlate with insurance success" and "measurable areas of personality"). However, he also believes that the sample of insurance salesmen employed in this study was singularly ill-chosen and has characteristics which serve to vitiate a number of the study's conclusions. Considerable work has been done in this field (2, 3, 4, 5, 6) and more is in progress. It is therefore important that major findings not be obscured by conclusions drawn from fragmentary and inadequate data.

Kahn and Hadley studied 84 "new life insurance agents" who had attended the Purdue Course in Life Insurance Marketing. It is implied that these men were a random group of individuals who had just entered the life insurance business. This, unfortunately, is not true. Many of these men had sold insurance before coming to the school. Furthermore, there is reason to believe that the companies and agency managers involved tended to send to the school those men whom they regarded as most promising. This seems probable in light of the fact that many of the men were subsidized to some degree by companies or managers during the course. Certainly, there is little evidence that the group is in any sense representative of new life insurance salesmen in general.

The authors state that the salesmen represented 19 life insurance companies. What they neglect to mention is that life insurance companies are not homogeneous with respect to agents' production. One study (5) has shown that, among 11 large insurance companies in Canada, the companies' median average monthly production of agents who survived 12 months ranged from \$5,500 to \$13,700 in the first year. Even among 7 United States companies of equivalent size, an analysis of variance shows that the first-year sales production of agents who survive for 12 months is heterogeneous at the 1% level.

In short, the sample employed is not relevant to the problems as stated, is curtailed to an unknown degree, and the criterion of success (sales for the duration of the school term) is contaminated by unrecognized and, with the number of salesmen involved, undetectable company differences.

Most of the conclusions listed by the authors are therefore questionable. It is stated that the correlation between sales during the first 13 weeks of selling and second period of 13 or more weeks is +.55. The statement should read during the first 13 weeks of selling *after entrance in a school.* It should be qualified by noting that the distribution is curtailed and that the correlation has probably been increased spuriously because of the effect of company differences.

The curtailment involved in the selection of the sample must also be considered in interpreting the statement that only one of the four personal history items investigated differentiates significantly between successful and unsuccessful life insurance salesmen. If the authors had employed more cases drawn from a sample of truly "new" agents and avoided widespread criterion categories, they would have found that age at entry has a significant, but curvilinear, relation to a success criterion (5, 6) and that minimum monthly income required has a similarly curvilinear and significant relation (6). They would also have found that agents with no dependents are significantly inferior to others in their first-year performance (6).

The conclusions concerning the differentiation of the criterion groups by various test items and total test scores is, of course, open to the same criticisms. Furthermore, the implication that this work is of value in the "identification of individuals for whom the likelihood of success is known" and, therefore, in the selection of life insurance agents, becomes highly suspect if it is remembered that many of these individuals were tested when their life insurance careers were well under way. The fact of success or failure may be a powerful determiner of test responses.

The problems of sampling, of restriction of range, and of criterion contamination are as real in investigations of the salesman as in any other.

Studies in which these problems are unrecognized or ignored can serve only to introduce further inaccuracies into an already confused field.

References

1. Kahn, D. F., and Hadley, J. M. Factors related to life insurance selling. *J. appl. Psychol.*, 1949, **33**, 132–140.

2. Life Insurance Agency Management Association. *2300 recruits a year later.* Hartford, Conn.: Life Insurance Agency Management Assoc., 1948, pp. 33.

3. ——. *Financing, survival, and production.* Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 12.

4. ——. *New agent characteristics.* Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 12.

5. ——. *Canadian recruiting and results.* Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 57.

6. ——. *Recruiting results.* Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 57.

A Comment on Wallace's Note on
"Factors Related to Life Insurance Selling"

J. M. Hadley and D. F. Kahn

Division of Education and Applied Psychology, Purdue University

It would appear that Wallace (9) is quite concerned that readers will misinterpret a recent article by Kahn and Hadley (3). Careful examination of the article in question will reveal that no generalizations are offered. The opening paragraph of the section entitled "Summary and Findings" on page 138 reads as follows: "Based solely on the criterion of written business, and pertaining only to those particular life insurance salesmen investigated in this study, the following conclusions may be drawn." All references in this section are to differences which "were found" to exist within the group of salesmen studied. No predictions were made concerning results which might be obtained from other samples. It is difficult for the writers to understand how "inaccuracies can be introduced into an already confused field" if research reports are read objectively and unintended generalizations are not inferred from admittedly "fragmentary and inadequate data."

Several of Wallace's points will be considered separately:

1. Wallace believes that "the sample of insurance salesmen employed in this study was singularly ill-chosen and has characteristics which serve to vitiate a number of the study's conclusions". The sample may be inadequate in many ways. It would be excellent if an entirely unselected sample could be obtained. It is doubted if such an entirely unselected sample was ever studied. The samples of recruits considered in the excellent studies by the Life Insurance Agency Management Association (4, 5, 6, 7, 8) are undoubtedly more adequate than those studied by Kahn and Hadley. Certainly, interest in being recruited also biases the samples studied by the Association to an unidentified degree. However, it is maintained that inadequacies inherent in the sample do not vitiate the conclusions concerning differences within the group.

2. Wallace criticizes the designation of the subjects of the study as "new life insurance agents." He also states that "many of these individuals were tested when their life insurance courses were well under way." A careful recheck of the data indicates that 95 per cent of the sample had not sold insurance before coming to the Purdue Life Insurance Marketing School. Actually, only four of the original 84 beginning

students had *ever* sold insurance. Two subjects had sold, or attempted to sell for longer than three months: one for nine months, and one for two years. The experimenters did not intend to include any subjects reporting more than three months' experience. Apparently two subjects were included by error. Shortly after the data were collected, the school began to require a minimum amount of experience. This was not true at the time the study was conducted. Furthermore, the original intent was to gather information of value to the school. In line with that purpose, it is submitted that the best sample would be classes of students in that school. Consequently, the new agents in classes I and II were selected. It is agreed that the sample is pre-selected by their companies and agency managers. The subjects may not be "in any sense representative of new life insurance salesmen in general" but they are representative of the first two classes attending the school. Again it is emphasized that the conclusions are limited to this group.

3. It is unfortunate that Kahn and Hadley in their condensed published article neglected to recognize the lack of homogeneity between life insurance companies and other complexities of the problem. Kahn (1, 2) in the original thesis has discussed the complexity of the problem at length.

4. Wallace states "most of the conclusions listed by the authors are therefore questionable." For some of the reasons which he states generalizations would be questionable, but one cannot question conclusions and results of a specific research study without questioning the integrity of the research workers. The writers accept the suggestion that the statement on page 135, line 7, of the results should read, ". . . during the first 13 weeks of selling after entrance in *the* school." It should be noted that the word "a" as suggested by Wallace has been changed to "the" by the writers. It would be interesting from the standpoint of research methodology to discover whether the effect of curtailment on the distribution and the effect of company differences, as discussed by Wallace, would increase or decrease the size of the reported coefficient of correlation.

5. Wallace seems to be disturbed that Kahn and Hadley did not obtain the same results as were obtained in several studies which he has reported. With a larger sample it is entirely possible that many differences would have been found to be more highly significant. On page 136 it is reported that age at entry, number of dependents, and minimum living expenses per month showed a positive relationship to the criterion. Dichotomies made in the range of each of the above-mentioned personal history items offer a means of showing the relationships between these measures and the criterion. Thirty-one agents of 30 years of age and

above averaged a mean weekly production of $5905 as contrasted with an average mean weekly production of $4612 for the 47 agents at age 29 or below. A similar average for 32 agents claiming two or more dependents was $6303 per week in comparison with $4307 per week for 46 agents claiming one or no dependents. For 13 agents requiring $280 a month or more as a minimum living expense the average mean weekly production was found to be $6554 as contrasted with $5022 for 59 agents requiring below $250 for living needs. Further manipulation of the data was not attempted because of the recognized inadequacy of the sample. Apparently the results obtained do tend to confirm those discussed by Wallace.

6. Wallace's criticism of the conclusions concerning the differentiation of the criterion groups by various test items and total test scores are, as previously discussed, again not considered relevant to the results but would be relevant to generalizations based on them.

7. Finally, Wallace states that the fact of success or failure may be a powerful determiner of test responses. This is granted, particularly in reference to the preference tests and to a lesser degree, personality tests. It is doubtful if it affects intelligence or biographical data. However, if test responses at any stated level of experience have any predictive value for future success, then they have validity for that purpose. In this connection some of the results of the study in question are indicative of the need for further research.

The writers gather from the general tone of the note that Wallace feels research in this area has been retarded and confused rather than advanced by the publication of the study being discussed. It is doubtful whether any research problem can be clarified by withholding legitimate data (and all data are legitimate) even though the population from which the data are derived is inadequate. Even single case datum is sometimes valuable. We must use care not to generalize beyond the scope of the data.

The writers would like to take this opportunity of urging that the valuable research by workers in the life insurance field be published in the scientific psychological journals so that it will be more readily available to academic research workers.

References

1. Kahn, D. F. *An analysis of life insurance salesmen.* Unpublished Master's thesis, Purdue University Libraries, West Lafayette, Indiana, 1946.
2. Kahn, D. F. *An analysis of factors related to life insurance selling.* Unpublished Doctorate Dissertation, Purdue University Libraries, West Lafayette, Indiana, 1948.

3. Kahn, D. F., and Hadley, J. M.   Factors related to life insurance selling.   *J. appl. Psychol.*, 1949, 33, 132–140.

4. Life Insurance Agency Management Association.   *2300 recruits a year later.*   Hartford, Conn.: Life Insurance Agency Management Assoc., 1948, pp. 33.

5. ——.   *Financing, survival, and production.*   Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 12.

6. ——.   *New agent characteristics.*   Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 12.

7. ——.   *Canadian recruiting and results.*   Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 57.

8. ——.   *Recruiting results.*   Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 57.

9. Wallace, S. Rains.   A note on Kahn and Hadley's "Factors Related to Life Insurance Selling."   *J. appl. Psychol.*, 1949, 33, —— – ——.

Instrument Reading. I. The Design of Long-Scale Indicators for Speed and Accuracy of Quantitative Readings *

Walter F. Grether

Aero Medical Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio

Quite a number of instruments used in aviation and elsewhere must be read with precision greater than can be provided by one revolution of a pointer on a circular dial of conventional size. There is considerable accumulated evidence that, except for the direct reading counter, most of the devices that have been used to increase effective scale length result in instruments that are very difficult to read. In a previous study by the author (2) on the design of clock dials, it was found that as common an instrument as a clock is quite difficult to read. Even the best clock designs required approximately 5 seconds (including recording time) for readings in hours and minutes by Air Force pilots. Even with this time spent on each reading, about 7 per cent of the readings on the better clocks were in error.

Aside from such laboratory data there is considerable evidence of instrument reading difficulties in the practical situations where these instruments are used. In a study of actual errors made by pilots in reading aircraft instruments carried out by Fitts and Jones (1), multiple-pointer or long-scale instruments provided the greatest number of serious cases of instrument misreading. The instrument reported as being misread most frequently was the altimeter. In the typical report the altimeter was read too high by a complete revolution of the most sensitive pointer, that is by 1000 feet. A tachometer designed with a rotating sub-dial to indicate RPM in thousands was likewise read too high by 1000 RPM. Numerous fatal and non-fatal accidents have been attributed directly to such instrument reading errors, and without doubt many of the unexplained crashes resulted from similar human failures.

The major purpose of the present investigation was to make a direct comparison in terms of speed and accuracy of quantitative readings of several of the possible methods of obtaining increased scale length on instruments. The experiment also had a secondary but more specific and practical purpose of finding improved methods of indicating altitude

in aircraft. For this reason all of the instruments were designed to read altitude in feet and all readings were made in feet as units.

It is emphasized that the evaluation of the different indicator designs in this investigation was with respect to the speed and accuracy of quantitative readings. Actually this is only one of several criteria which most instruments should be required to satisfy. It has been pointed out by the author (3) that in aviation in particular there would appear to be at least three major ways in which instruments may be read, depending upon the purpose of the reader. These three types of reading may be categorized as follows: a. Check reading—for assurance of a null, normal, or desired indication; b. Qualitative reading—for the direction and approximate magnitude of a deviation from a null, normal, or desired indication; and c. Quantitative reading—for the numerical value of an indication.

The above categories of instrument reading have considerable utility as criteria against which to evaluate different instrument designs. It is usually possible from a knowledge of the situation in which an instrument is to be used to decide the reading purposes or criteria which it is most necessary to satisfy. The criteria against which an instrument is to be evaluated then provide operational definitions of the experimental measurements to be made. As mentioned earlier, the experimental indicators in this investigation were evaluated only with respect to the third criterion, quantitative reading. In this study, furthermore, there was no concern with small errors of interpolation, only with larger errors resulting from assignment of incorrect values to graduation marks.

Experimental Procedure

Nine experimental altitude indicator designs were used in this investigation. These are shown along with some of the results in Figure 1. The first of these indicators, design A, is a simulation of the altimeter almost universally used in military and larger commercial aircraft. On this instrument the longest pointer gives readings in hundreds of feet, the broad pointer is read on the same scale in thousands of feet, and the small pointer is read on the same scale in ten-thousands of feet. Altimeter designs B and C also simulate existing but not commonly used types.

Altimeter design D uses a single pointer to indicate altitude in hundreds of feet. This pointer makes one revolution for each 1000 feet change in altitude and the multiples of 1000 feet are indicated on a simulated direct reading counter. This counter has two drums, one for 1000-foot and the other for 10,000-foot increments. It is assumed that the motion of these drums would be intermittent and that single whole numbers would always be showing.

In design E, also, only one pointer is used, but two dials rotating behind a window indicate the multiples of 1000 feet. In this design the motion of the dials showing through the window is assumed to be continuous rather than intermittent, thus permitting more than one number (or half numbers) to appear.

Design F indicates altitude in quite a different manner from the other

instruments. In this display the pointer is assumed to make only one revolution to cover the entire altitude range. The range being covered is indicated in the window as 0–1000 feet, 0–10,000 feet, or 0–100,000 feet. The meaning of the numerals on the dial graduations is, therefore, determined by the range indicated in the window. This indicator is similar in principle to a radio altimeter now in use. It is obvious that the precision of indication on such an instrument decreases as the range being covered increases.

Altimeter designs G and H are similar in that they simulate a scale moving vertically behind a window. An instrument following design G could use either an endless tape or drum to present the moving scale, with a counter to indicate multiples of 1000 feet. An instrument using design H would require a very long tape with a scale covering the desired altitude range.
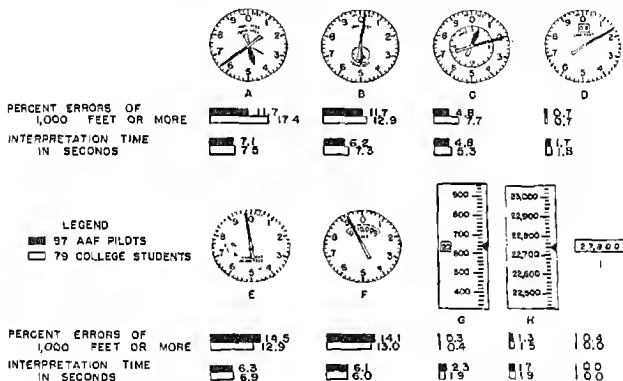
Fig. 1. Speed and accuracy in reading altitude from different types of instruments.

The last experimental design, I, simulates a simple direct reading counter without any moving pointer or scale. For reasons pointed out later in the discussion of results, such an indicator would probably be unsatisfactory for the pilot, but might be suitable for other aircrew members such as the navigator. One of the major reasons for including it in this study was to get an approximate measure of the time required to copy a series of numbers representing an altitude reading, it being assumed that no interpretation time would be involved in reading altitude from this type of indicator.

For each of the altimeter designs used in this experiment a test booklet was prepared. The cover (page 1) of each booklet presented the experimental subject with detailed instructions for reading the dial design in that booklet, and a sample dial for the subject to read. On the two inside pages, 2 and 3, the dial design was reproduced with 12 different settings. Under each picture was a space for writing in the reading.[1]

[1] The large number of drawings needed for the nine test booklets were produced by Miss Mary Cowles of the Psychology Branch with the photographic assistance of Mr. D. M. Penrose of the Laboratory Services Unit of the Aero Medical Laboratory.

Special precautions were taken in the preparation of the drawings and choice of altitude settings to be used in the various test booklets to prevent biasing the results for or against any of the indicator designs. The circular dials were 2¼ inches in diameter. From this other dimensions can be estimated from Figure 1. All essential numerals and graduation marks were sufficiently large and distinct to be easily legible. Except for the inner dials on designs B and C all scales were alike in having numerals at all 100-foot graduations with intervening marks at 20-foot intervals. Other factors equalized were the number of settings above and below 10,000 feet, the number of sensitive pointer settings on 100-foot graduation marks, the number of sensitive pointer settings just preceding and just following the zero on the scale, and the number of sensitive pointer settings on the left and right halves of the dial. Precautions were also taken to be sure that no essential information was hidden by any of the hands, and that the interrelationships between pointer positions were correct. For indicator design F some of the settings were midway between graduation marks. For the remaining designs the sensitive pointer (or reference mark) was always on a graduation mark. Thus, no interpolation was required to obtain correct readings.

The altimeter reading test was taken by 97 USAF pilots in the Instrument School at Barksdale Field, Louisiana, and 79 college men (without aircrew experience) at Denison University, Granville, Ohio. In administering the test, the booklet for only one altimeter design was passed out at a time, and sufficient time was allowed for reading the instructions and working the sample item. At a signal all subjects opened the booklet and worked until completing all items. Each subject's completion time was recorded on his booklet. Four sequences for administering the nine test booklets were used in order to counterbalance for learning effects. An approximately equal number of subjects (in each of the two subject groups) took the test in each sequence.

The two subject groups of dissimilar experience were used in order to get some measure of the effect of experience on the ability to read the various dial designs. All of the USAF pilots can be assumed to have spent several years flying with altimeter design A, and possibly some experience with designs B and C. The college men can be assumed to have had little, if any, experience in reading altitude from any type of indicator. In general intelligence and education the two groups were very similar.

Results of Comparisons Among Indicator Designs

The data obtained in this investigation were analyzed to determine the frequency of errors and the time per instrument reading. These results are shown in Table I which gives the per cent of total readings in error for the nine indicator designs.[2] None of the errors included in this table resulted from inaccuracies in pointer interpolation since all settings of the sensitive pointers were on graduation marks (except for design F which had some settings midway between marks).

Data on speed of reading are also shown in Table 1. It will be recalled that the subjects wrote their answers in the test booklets and the time for completion was recorded in each instance. The average time per reading could thus be computed from the total time and the total number

---

[2] Altimeter reading errors during actual flight probably occur with much lower frequency than found in this study, since in flight the pilot can anticipate the approximate readings.

Table 1

Altitude Indicator Design	USAF Pilots, N = 97		College Men, N = 79	
	Per Cent Errors (a)	Seconds per Reading* (b)	Per Cent Errors (c)	Seconds per Reading* (d)
A	15.9	9.6	20.8	9.8
B	15.0	8.6†	17.9	9.6
C	8.3†	7.3†	11.4†	7.8†
D	3.5†	4.2†	2.1†	4.1†
E	17.3	8.8	15.3†	9.2
F	24.1	8.7†	21.0	8.3†
G	2.1†	4.8†	3.0†	4.2†
H	2.5†	4.2†	4.5†	4.2†
I	0.6†	2.5†	0.3†	2.3†

	N	r	Confidence level
Correlation between speed and accuracy for different designs:			
For pilots (columns a and b)	9 designs	.91	1%
For college students (columns c and d)	9 designs	.95	1%
Correlation between pilots and college students on different designs:			
Per cent errors (columns a and c)	9 designs	.95	1%
Seconds per reading (columns b and d)	9 designs	.99	1%
Correlation between speed and accuracy of individuals on all designs:			
Pilots	97 pilots	.38	1%
College students	79 college students	.44	1%

* Reading time included time for subject himself to record answer.
† Indicates statistical significance (at one per cent level of confidence) of superiority over conventional altimeter (design A).

of items, but this time included the time for recording as well as for reading or interpreting the instrument.

A reproduction of each of the experimental indicator designs accompanied by graphic illustrations of the more significant findings is provided in Figure 1. The upper pair of bars under each indicator shows the per cent of errors equal to or exceeding 1000 feet for the two groups of subjects. The lower pair of bars gives the computed interpretation time for each of the two groups of subjects. An estimate of the time for interpretation only was obtained by subtracting from the average time for each design the average time for design I (the direct reading counter). The reading of altitude from design I involved the mere copying of the numbers shown, and hence was assumed to require no interpretation.

Discussion of Results

Indicator Designs A, B, and C. The results of this investigation, as shown in Figure 1 and Table 1, show that Design A, which simulates the conventional altimeter, is a very difficult instrument from which to obtain quantitative readings as required in this study. Even the pilots, all of whom had spent several years flying with this instrument, spent more time per reading on this indicator than on any of the other designs studied. Only one of the remaining eight indicators, design F, resulted in a higher proportion of errors. It must be concluded that it is a very difficult task to combine into a single numerical value the readings of three pointers indicating on a single scale, as required in reading the conventional altimeter. Designs B and C apparently were slightly easier to read than design A.

Indicator Design D.[3] This indicator uses only one pointer, with the 1000-foot and 10,000-foot indications provided by a counter. Such a combination proved to be very easy to read. For USAF pilots the per cent of total errors was very low, 3.5 per cent, and only 1.7 sec. was required for interpretation (as contrasted with 15.9 per cent and 7.1 sec. for the conventional altimeter). More significant, perhaps, is the finding that only 0.7 per cent of the readings erred by more than 1000 feet. Most of the errors in reading indicator design D resulted from assigning 10 feet instead of 20 feet to each of the graduation intervals between numerals.

Indicator Design E. The substitution for two of the pointers on the altimeter of two rotating dials appearing through a window appears to have no advantage. This indicator was designed so that under most circumstances only one number would appear on each of the two rotating dials. But if such dials rotate continuously (rather than intermittently) during altitude changes, as assumed in this test, it is inevitable that at certain settings two numbers will be equally visible. Such indications are very difficult to read correctly.

Indicator Design F. On this indicator the range covered by the indicating pointer and scale is dependent upon range limits shown in the window. The high proportion of errors and slow reading time suggest that the required changes in interpretation of the primary scale are difficult for human beings to carry out.

---

[3] On the basis of this study indicator design D, combining a sensitive pointer with a direct reading counter, was recommended as a replacement for the existing three-pointer altimeter. As a consequence the Kollsman Instrument Division of the Square D Company is now developing such an altimeter. Two other items of aviation equipment currently being developed by the Air Force, an absolute (radio) altimeter and an airborne distance measuring device, are also incorporating this type of indication.

Table 2

Frequency of Various Types of Error Made by 97 USAF Pilots and 79 College Students
in Reading the Conventional Three-Pointer Altimeter

Description of Error		Per Cent of Total Readings in Which Error Appeared	
		Pilots	College Students
Reading to nearest numeral instead of to lower adjacent numeral—			
(Failure to consider more sensitive pointer)	100 Ft.	0.09	0.11
	1,000 Ft.	2.58	1.48
	10,000 Ft.	1.72	2.11
	Total	4.39	3.69
Reading to lower adjacent numeral when nearest numeral is correct—			
(Failure to consider more sensitive pointer)	100 Ft.	0.0	0.0
	1,000 Ft.	0.26	2.22
	10,000 Ft.	0.0	0.11
	Total	0.26	2.32
Displacement of digit in number series—			
(Interchange of digit with adjacent zero)	20 Ft.	0.17	0.42
	100 Ft.	0.86	0.95
	1,000 Ft.	2.06	2.64
	10,000 Ft.	0.86	1.48
	Total	3.95	5.48
Misreading of scale or numeral—			
	20 Ft.	3.09	2.64
	100 Ft.	1.20	1.05
	1,000 Ft.	1.46	2.85
	10,000 Ft.	0.09	0.53
	Total	5.84	7.07
Omission of one pointer—			
	100 Ft.	0.0	0.0
	1,000 Ft.	0.86	0.21
	10,000 Ft.	0.86	1.05
	Total	1.72	1.27
Pointer exchange—			
	100 and 1,000 Ft.	0.17	0.84
	100 and 10,000 Ft.	0.0	0.0
	1,000 and 10,000 Ft.	0.09	1.48
	Total	0.26	2.32
Repetition of reading on one pointer—		0.95	0.84
Complex and unclassified errors		0.86	1.48

Indicator Designs G and H. The vertical moving scale instruments proved to be easy to read in this experiment. The virtues of such instruments for check reading and qualitative reading were not evaluated in this study.

Indicator Design I. This indicator, which simulates a simple Veeder counter, was read with greatest speed and accuracy of all the indicators. This would suggest that where only quantitative readings are to be provided this would be the most desirable type of instrument. It is believed that for check reading and for qualitative reading such an instrument would be inferior to one using a moving pointer.

Types of Error in Reading the Three-Pointer Altimeter

Because of the widespread use of the three-pointer altimeter, and because of the frequent use of this type of multiple pointer indication for other purposes, it seemed worth-while to make a more detailed analysis of the types of errors made in quantitative readings of this instrument. This analysis was based on the same data that have already been summarized in Table 1 and Figure 1. It will be recalled that 97 USAF pilots and 79 college students each made 12 readings on the three-pointer altimeter. This gave a total of 1164 readings by pilots and 948 by non-pilots.

The detailed classification of errors into types and sub-types is shown in Table 2 along with the per cent of total readings in which each occurred. Two or more types or sub-types of errors were in some cases charged against a single erroneous reading. For this reason the figures in the per cent columns total up to more than the total per cent errors as reported in Table 1. For detailed descriptions of all the error types, and the assumed mental processes which led to the incorrect answers, the reader is referred to Aero Medical Laboratory Memorandum Report No. MCREXD-694-14A.

Discussion

In an experiment such as this a number of questions arise with regard to the suitability of the criterion measures which have been used and with regard to the effect of the subject group upon the results. For this reason there have been included in Table 1 a number of correlation coefficients which bear on these problems.

A serious question facing the experimenter is the effect of the experience of the subject group upon the validity of the findings. In the present experiment two subject groups were used which represented extremes in experience as related to the task being performed. All USAF pilots had had considerable experience in reading one of the experimental indicator designs along with general experience in reading aircraft instru-

ments. The college students, on the other hand, included no pilots or other military air crew members. In spite of this difference in background of experience the two groups gave highly similar results as indicated by a correlation between the two groups of .95 on per cent errors and .99 on seconds per reading. This would suggest that the previous experience of the subjects is of relatively minor importance in an experiment of this type.

In the present experiment neither speed nor accuracy of response were controlled, thus making possible two independent criteria for evaluation of the different dial designs. In Table 1 the correlations between speed and accuracy for the different dial designs are .91 for pilots and .95 for college students, indicating very high agreement between the two criteria for goodness of the several indicator designs. Or stated differently, the indicator designs which were read most rapidly were also read most accurately. Correlation coefficients between speed and accuracy of individuals for all designs are also positive, but much lower, .38 for pilots and .44 for college students. These values indicate, however, that in general the individuals who read the indicators most rapidly also read them most accurately. In a previous study by the author (2) on clock dial designs the correlation coefficients were likewise positive, but somewhat lower in magnitude.

In two previous experiments on instrument design by Loucks (4) and Sleight (5) a somewhat different technique was used in that the instrument exposure time was controlled tachistoscopically and only accuracy of reading was measured. Such a technique might be expected to force an increased error rate and thus accentuate the differences between indicator designs. It is the belief of the author, however, that such a control of exposure does not constitute control over response time, but serves rather to restrict the number of visual fixations of the displayed material. The actual response may be delayed for several seconds during which the subject retains a mental image of the indicator scale and pointer.

It is quite possible that in the experiment of Sleight (5) the use of a controlled exposure time which did not permit a change in the preparatory eye fixation led to erroneous findings. It is believed that this technique favored the fixed pointer indicators on which the subject was able to anticipate the location of the pointer. The two fixed pointer indicators in the present study, designs G and H, showed no general superiority over the only comparable moving pointer indicator, design D.

Summary

An evaluation was made of the speed and accuracy with which quantitative readings could be made of nine experimental altitude indi-

cators. The results are considered to apply also to other types of quantitative indication which require very great scale length. Evaluation of the various indicator designs was made by having 97 USAF pilots and 79 college men read 12 settings on each instrument. The instrument faces were reproduced in test booklets which provided spaces for writing in the readings. Both accuracy and speed-of-reading data were obtained for each of the nine indicator designs.

The major conclusions indicated by the results of this investigation are as follows:

1. The combining into a single numerical value of the indications from two or more pointers, or from a pointer and rotating subdials, is a relatively difficult task for human beings. Such instruments are conducive to very large errors in reading.

2. The ease with which long scale indicators can be read quantitatively appears to depend upon the extent to which the digits are already combined in the proper sequence by the instrument.

3. A multiple pointer instrument such as the altimeter with continuous motion of the non-sensitive pointers ie frequently read too high by a complete revolution of the sensitive pointer.

4. The speed and accuracy of instrument reading are positively correlated, indicating that gains in reading speed can normally be expected to improve accuracy also.

5. College men without altimeter reading experience showed virtually the same pattern of results in this study as highly experienced USAF pilots, suggesting that instrument reading difficulties are quite basic in nature and not readily modified by experience.

References

1. Fitts, P. M. and Jones, R. E. Psychological aspects of instrument display. I. Analysis of 270 "pilot error" experiences in reading and interpreting aircraft instruments. USAF Air Materiel Command Memorandum Report No. TSEAA-694-12A, 1947.

2. Grether, Walter F. Factors in the design of clock dials which affect speed and accuracy of readings in the 2400-hour time system. *J. appl. Psychol.*, 1948, 32, 159–169.

3. Grether, Walter F. Designing instrument dials for quick, accurate reading. Machine Design, 1948, 20, 150–152 and 208–209.

4. Loucks, R. B. Legibility of aircraft instrument dials: The relative legibility of tachometer dials. AAF School of Aviation Medicine, Project No. 265, Report No. 1, 1944.

5. Sleight, Robert B. The effect of instrument dial shape on legibility. *J. appl. Psychol.*, 1948, 32, 170–188.

Types of Errors in Location Judgments on Scaled Surfaces.
I. Errors of Configuration *

Adelbert Ford

Department of Psychology, Lehigh University

Many instruments have been so designed that an operator is required to locate the position of a signal on a flat area with reference to a superimposed system of scales, which may be either polar or rectangular. This study deals with the latter. The nature of this signal may be a small dot, or a white patch of small dimensions, which appears suddenly and must be reported for its elevation on the y-axis and its horizontal location on the x-axis. This is typical in using some of the types of radar cathode-ray scopes.

In practice there are two systems for keeping such a signal or spot located. 1. A transparent plastic scale, engraved with suitable reference lines, may be placed over the area, the operator locates the proper line, follows to the end, and notes the position between engraved numerals, interpolating between lines when necessary. 2. The operator may be provided with a "tracker" which he pushes around the face of the area, keeping it superimposed on the signal, and this mechanism registers the x-and y-values on remotely located meters. The latter has been found to be objectionable because it usually required two operators, and it is mechanically complex. However, the simple method of using scaling assistance may involve intolerable errors, greater mental concentration, and therefore it is desirable to know just what kinds of errors an average operator does commit in using scaling assistance, on the basis of quantitative experimentation. If these types of errors are found to be intolerable, then it is worth while to pursue such engineering design as may eliminate the human error in scale reading methods.

The present study deals with the first of four types of reading errors

on scaled surfaces. This will be called errors of *configuration* because we shall show that the configuration or *shape* of the field produces systematic errors in one part of the field as contrasted with errors in another part of the field.

A sector scope is essentially a triangular area. This sets the condition for the *perspective illusion.* Objects near the apex of the triangular space appear larger than at the open end of the triangle. It would be expected that elevation judgments, with respect to a zero line of reference would be correspondingly exaggerated. The questions are: How much? Are all people susceptible?

There are many citations of general principle in the literature. Ponzo (1) showed the principle with respect to estimated lengths of lines. Köhler and Wallach (2) maintained that space estimations at the open end were underestimated while those at the apex were over-estimated.

Method

1. *Artificial Series.* The types of scope faces used in the artificial series are presented in Figures 1 to 6. The figures on the left are for the sector-type of radar scope, commonly used, and show the condition for the perspective illusion. The figures on the right are rectangular pres-entations used as a "control" with the same kind of problems. All scope pictures were 7 inches in diameter, viewed at a distance of 16 inches, or (in group experiments) with an equivalent visual angle.

2. *Natural Series.* Figure 7 exhibits a photograph of a real radar sector-type scope, one of the stimulus series which we presented with the ultra-violet radar simulator. The white spot at the right is a signal from
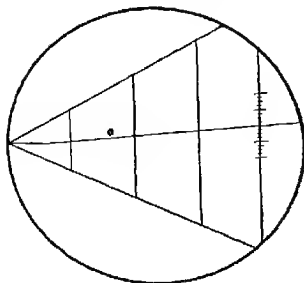
FIG. 1. Sector "unscaled" scope.   **FIG. 2.** Rectangular "unscaled" scope.
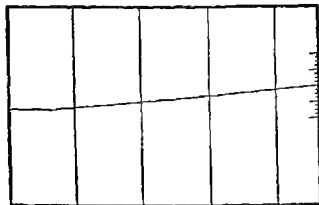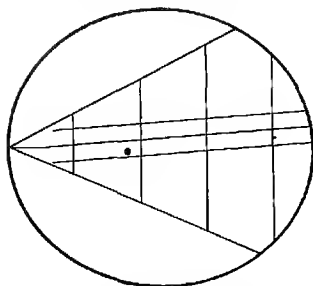
FIG. 3. Sector scope with 100-foot side lines.
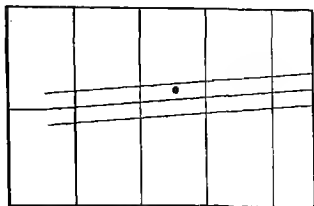
FIG. 4. Rectangular scope with 100-foot side lines.
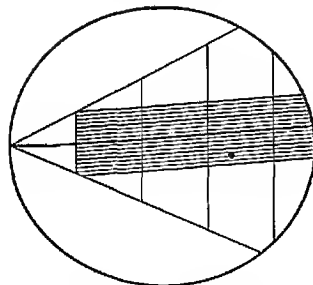
FIG. 5. Sector scope with multiple scaling.
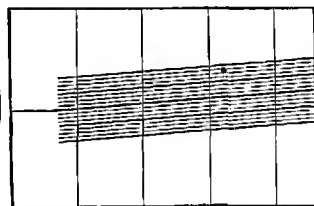
FIG. 6. Rectangular scope with multiple scaling.

an approaching airplane about to land. In this series it was impossible to use the rectangular scope for comparison, because no scopes were made that way.

All projected images were on a phosphorescent radar screen, of typical color and signal persistence, except in the group experiments. Signals were presented serially in a fairly realistic rate of progression. All signals were white on a black field.

Subjects were scored for error and reaction time by recording verbal answers as rapidly as made, and using a chronoscope actuated by a voice key.

There were three types of scaling assistance, as indicated in Figures
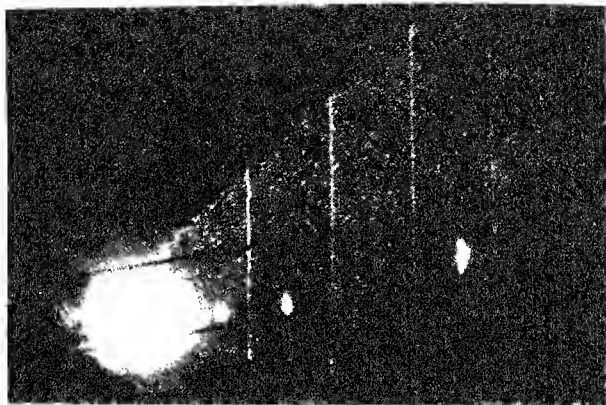
Fig. 7. Photograph of sector presentation on a real
radar scope used for ground-control-approach.

1 to 6. (a) The "unscaled" scope showed merely a zero line of reference
with a marginal standard for space values (Figures 1 and 2). (b) The
"100-foot scope" used the same zero reference line, but had a parallel
line, above and below, to show 100 feet of signal deflection. The lines
representing 100 feet were 0.4 inch above and below the zero line (Figures
3 and 4). (c) The scope with "multiple scaling" had fine lines, 0.1 inch
apart, with every fourth line heavier. Fine lines represented increments
of 25 feet. Heavy lines represented 100 feet deflection increments
(Figures 5 and 6).

Results [1]

It will be necessary to remember that these experiments were run on
rather complex equipment, generally one run at a time, and that the
number of readings possible was thus restricted, and the number of
subjects used was necessarily limited. All differences were subjected to
calculations of critical ratio, and the differences in means with a signifi-
cance better than the one percent level have been shown in italics in all
tables. It will be obvious that statements made under *conclusions* are
qualified.

It was not possible to keep subjects completely unsophisticated with respect to the existence of the illusion, during the 18 months of work. Therefore we shall show training effects by separate tables.

1.   *Untrained Subjects on Artificial Scopes.*   The introductory or trial experiments were done with the artificial scope pictures of Figures 1 to 6.   The trend or "drift" of possible overestimation compares the one-third area of the open end of the sector with the one-third area at the apex end of the scope field.   This trend is reduced to a single figure expressed in percentage of over or underestimation.   A similar problem is then shown for the same subject on a rectangular scope, calculated in the same manner, and used as a control.   Presentations were randomized.

Table 1

The Perspective Illusion in Untrained Subjects
Percentages of Error in Elevation Judgments
Individual Experiments, Artificial Scope
Combined Data for Four Subjects
Plus signs indicate over estimations; minus signs, under estimations.

Type of Scaling Assistance	Sector Scope (Experimental)			Rectangular Scope (Control)			Difference (Illusory Trend)
	Left (Apex)	Center	Right (Open End)	Left	Center	Right	
Unscaled	+19.6	−9.8	−.01	−1.1	−.3	+2.3	*+23.1*
100 ft. Lines	+14.0	−.3	+2.4	+.8	+3.1	−.4	*+10.4*
Multiple Lines	−4.5	+2.8	+1.1	+.1	+2.5	+1.5	−4.2

Note: Difference values expressed in italics are better than the one per cent level of significance.   Values are averages based on 40 to 55 runs.

The first, or introductory runs, showed a strong illusory trend of overestimation at the apex of the sector scope (Table 1), when the field was "unscaled."   The introduction of a pair of 100-foot reference lines (0.4 inch on each side of the line of zero position) had a marked effect in reducing the illusion, but did not eliminate it.   When the multiple scaling system was used (Figures 3 and 6) there was no reliable evidence of illusory effect.

2.   *Partially Trained Subjects on Artificial Scopes.*   After 40 to 55 runs, the next section of the training series (Table 2) showed a reduction in the amount of the illusion on the unscaled scope, and no significant illusory trends for the scopes with 100 foot scaling lines, or the multiple scaling system.   Evidently scaling reduces the illusion.

3.   *Individual Differences.*   Dealing with the spread of "random

Table 2

The Perspective Illusion, Second Training Stage
Percentages of Error in Elevation Judgments
Individual Experiments, Artificial Scope
Combined Data for Four Subjects

Plus signs indicate over estimations; minus signs, under estimations.

Type of Scaling Assistance	Sector Scope (Experimental)			Rectangular Scope (Control)			Difference (Illusory Trend)
	Left (Apex)	Center	Right (Open End)	Left	Center	Right	
Unscaled	+11.8	+6.9	−2.7	−2.0	−7.1	+.7	+17.2
100 ft. Lines	+3.9	+3.1	+4.1	−.3	+2.1	+.7	+.8
Multiple Lines	−.3	−1.8	+.6	+.8	+.4	+1.7	−1.8

Note: Difference values expressed in italics are better than the one per cent level of significance. Values are based on 55 to 109 runs.

errors" all subjects were very much alike, showing standard errors closely similar. However, with respect to the illusion, it had become evident that some subjects have strong susceptibility, and an occasional subject seems to be completely free from any proneness. Table 3 is presented to show individual differences. One subject entered too late to be given the runs on the unscaled scope, but is included to complete the data.

Table 3

The Perspective Illusion, Second Training Stage
Percentages of Error in Elevation Judgments
Individual Experiments, Artificial Scope
Data for Individual Differences

Plus signs indicate over estimations; minus signs, under estimations.

Initials of Subjects	Unscaled Scope			100-ft. Reference Lines			Multiple Scaling		
	Sector Scope (Experimental)	Rectangular (Control)	Difference (Illusory Trend)	Sector Scope (Experimental)	Rectangular (Control)	Difference (Illusory Trend)	Sector Scope (Experimental)	Rectangular (Control)	Difference (Illusory Trend)
R. J. R.	+27.1	−1.5	+28.6	+.3	−1.4	+1.7	−.9	+5.0	−5.9
F. P. H.	+16.9	−7.8	+24.7	+.6	−2.1	+2.7	+.3	−2.8	+3.1
L. A. A.	+12.3	−1.0	+13.3	+1.9	−.2	+2.1	−.4	−2.9	+2.5
D. M. S.	+3.2	+1.6	+.6	−9.0	+1.0	−10.0	−2.3	−.3	−2.0
W. A. S.	(New Subject)			+2.2	−5.5	+7.7	−1.7	−4.1	+2.4

Note: Difference values expressed in italics are better than the one per cent level of significance. Values are based on from 13 to 25 runs.

This is a breakdown from data in the series for partially trained subjects. Three subjects are quite strongly susceptible. One is not.

This led us into an experiment in which the same problems were projected on a large screen for a group experiment on untrained subjects, with visual angle kept approximately the same. Table 4 arranges these subjects in an order of most susceptible to least susceptible. The purpose of this section was to secure more extended data on individual

Table 4

The Perspective Illusion in Untrained Subjects

Percentages of Error in Elevation Judgments

Group Experiments, Artificial Scope

Plus signs indicate over estimations; minus signs, under estimations.

Initials of Subjects	Unscaled Scope			100-ft. Reference Lines			Multiple Scaling		
	Sector Scope (Experimental)	Rectangular (Control)	Difference (Illusory Trend)	Sector Scope (Experimental)	Rectangular (Control)	Difference (Illusory Trend)	Sector Scope (Experimental)	Rectangular (Control)	Difference (Illusory Trend)
N. J. R.	+33.3	−31.2	+64.5	+1.3	+7.4	−6.1	−2.4	+8.0	−10.4
L. E. K.	+55.5	−.6	+56.1	+8.3	−16.7	+25.0	−2.3	+5.9	−8.1
B. J. J.	+52.8	−2.9	+55.7	+39.3	+.8	+38.5	−7.4	−3.3	−4.1
W. J. K.	+52.8	−2.0	+54.8	+3.3	−4.3	+7.6	+16.8	−17.2	+34.0
D. L. H.	+49.8	−.8	+50.6	+.5	−2.3	+2.8	+6.7	−11.6	+18.3
M. R. K. S.	+65.0	+14.6	+50.4	+0.9	+2.6	+4.3	+3.6	+3.7	−.1
J. H. J.	+36.5	+.1	+36.4	+6.4	−5.8	+12.2	−.0	−.7	+.1
M. J. D.	+26.1	−5.4	+31.5	+5.6	−3.1	+8.7	−5.0	−2.2	−2.8
R. K. S.	+17.2	−11.1	+28.3	−3.1	−.9	−2.2	−2.4	+1.7	−1.1
R. B. T.	+25.5	+1.3	+24.2	−.1	+.2	−.3	+.7	−1.5	+2.2
M. B. C.	+18.2	−5.2	+23.4	+3.1	−1.5	+4.6	−2.4	+3.8	−6.2
A. W. R.	+14.8	−6.0	+20.8	+6.9	−1.3	+8.2	+.6	+2.5	−1.9
S. K. M.	+23.7	+3.1	+20.6	+3.1	−20.7	+23.8	+1.4	−1.0	+2.4
A. P. R.	+16.9	−3.7	+20.6	+.7	+.5	+.2	+23.3	−18.9	+42.2
C. A. W.	+31.4	+12.9	+18.5	−6.1	−9.9	+3.8	+1.5	−4.1	+5.6
A. H. F.	+11.4	−5.0	+16.4	−1.1	−2.6	+1.5	−3.1	−.8	−2.3
C. E. F.	+23.3	+7.5	+15.8	+23.3	−10.3	+33.6	−1.9	+.9	−2.8
M. S. W.	+5.8	+2.2	+3.6	+1.0	−16.0	+17.0	−.9	−10.3	+9.4
D. A. W.	−.4	+13.9	−14.3	+.5	−3.0	+3.5	+18.9	−1.2	+20.1

Note: Difference values expressed in italics are better than the one per cent. level of significance. Values are averages based on from 14 to 28 runs.

differences. Maximum susceptibility reached 65% overestimation at the apex of the sector sweep.

4. *Untrained Subjects on Reproductions of Field Scopes.* Six new and untrained subjects were now tried on reproductions of radar field runs, using the ultra-violet projection radar simulator. Three are strongly susceptible (see Figure 7 and Table 5). One is suspected of being moderately susceptible. Two showed no reliable differences.

Table 5

The Perspective Illusion in Untrained Subjects

Percentages of Error in Elevation Judgments

Individual Experiments, Reproductions from Field Scopes

Plus signs indicate over estimations; minus signs, under estimations.

Initials of Subjects	Unscaled Scope			100-ft. Reference Lines			Multiple Scaling		
	Apex of Sector	Open End	Difference (Illusory Trend)	Apex of Sector	Open End	Difference (Illusory Trend)	Apex of Sector	Open End	Difference (Illusory Trend)
J. H. J.	+21.4	−23.4	+44.8						
R. C.	+4.4	−21.9	+26.3	−1.0	−13.9	+12.9	+2.7	−14.6	+17.3
F. P. A.	+4.6	−17.1	+21.7				−5.8	−5.7	−.1
C. R. B.	−6.3	−13.8	+7.5	+8.7	−22.7	+31.4	−6.9	+.1	−7.0
J. H. F.	−7.0	−13.3	+6.3	−3.0	−5.5	+2.5	−.6	−1.6	+1.0
F. G. F.	−14.5	−2.8	−11.7	−2.7	−9.9	+7.2	−6.5	+7.0	−13.5

Note: Difference values expressed in italics are better than the one per cent level of significance. Values are averages based on from 17 to 48 runs.

5. *Trained Subjects on Reproductions of Field Scopes.* There were two subjects who had started with the program, 18 months before. These are considered as trained subjects. The data are presented in Table 6. Neither has a reliable difference between the apex and the open end of the scope, either for unscaled or scaled scopes. One subject, L. A. A., had been mildly susceptible in the earlier stages (see Table 3). The

Table 6

The Perspective Illusion in Trained Subjects

Percentages of Error in Elevation Judgments

Individual Experiments, Reproductions from Field Scopes

Plus signs indicate over estimations; minus signs, under estimations.

Initials of Subjects	Unscaled Scope			100-ft. Reference Lines			Multiple Scaling		
	Apex of Sector	Open End	Difference (Illusory Trend)	Apex of Sector	Open End	Difference (Illusory Trend)	Apex of Sector	Open End	Difference (Illusory Trend)
L. A. A.	−7.8	−9.0	+1.2	−6.0	+2.7	−8.7	−2.2	+3.9	−6.1
D. M. S.	−5.5	−1.5	−4.0	−5.6	−2.8	−2.8	−2.9	−2.7	−.3

Note: All differences have a significance poorer than the one per cent level.

other subject, D. M. S., had never been susceptible, even on the artificial scopes. There seems to be evidence that, for some subjects, training either decreases the illusory effect, or even eliminates it.

6. *The Köhler-Wallach Principle.* In dealing with the perspective illusion Köhler and Wallach (2) stated that space judgments in a triangular area showed *overestimation* at the apex and also *underestimation* at the open end, whereas many of the textbooks stress only the overestimation at the point of the triangle. The Köhler-Wallach principle is well illustrated among the first three subjects of Table 5. In the artificial scope series the principle was still there, though this has been omitted in the form of our tabulation, but it was much milder, with only slight tendencies toward underestimation at the open end.

Summary

1. When the position of signals on the area of a sector-type scope reaches the apex of the scan-line sweep, in what is essentially a triangular area, overestimations of space reach as much as 65% for some subjects.

2. The great majority of all subjects show some degree of susceptibility, but the range of individual differences extends from complete lack of proneness to about 65% relative overestimation at the apex. Within the limitations of the number of subjects used, it might be expected that 90% of all subjects show some degree of the illusion.

3. The design of the field of a radar scope must take into consideration the shape of the field for its total effect on scope reading errors. With respect to this illusion, clearly visible multiple scaling will reduce the space distortion, but judgment must be deferred lest other types of errors are introduced, and these will be presented in later articles.

References

1. Ponzo, M. Urteilstäuschungen über Mengen. *Archiv. für die Gesamte Psychologie*, 1928, 65, p. 135.
2. Köhler, W., and Wallach, H. Figural after-effects, an investigation of visual processes. *Proceedings of the American Philosophical Society*, Vol. 88, No. 4, October 1944, p. 288.

Types of Errors in Location Judgments on Scaled Surfaces. II. Random and Systematic Errors *

Adelbert Ford

Department of Psychology, Lehigh University

A large variety of instruments require operators to report the position of a "signal," such as a white spot, by reading its position with reference to superimposed scaling lines. In dealing with types of radar associated with the navigation of aircraft a single large error could cause loss of life and the destruction of expensive equipment.

In the last article[1] we noted the existence of errors caused by the shape of the field surface. In the present article, using the same scaling and problem sequences, we propose to show: (1) the size of the *random errors* caused by the limiting effects of interpolating scale values of specific scales, and (2) certain systematic errors consisting in particular of the *confusion error*, defined as a mistaken interpretation of the numerical value of the scale points, and what we shall call *persistence errors*, defined as a proneness of some subjects to bias reports in a sequential series by memory effects of the previous reports.

Although the present report is specifically concerned with position reporting from scaled areas, it will probably be instantly perceived that some of the principles are perhaps equally applicable to linear scales. The consequences of this error analysis are much more basic than the narrow application to radar scopes.

Fineness of Scaling and Random Error[2]

As illustrated in the previous article, there were three types of scaling used for these experiments: (1) a scope with a zero line of reference across the field, but no other scaling assistance other than a sample scale printed

on the side of the scope for comparison; (2) a scope with a so-called "100-foot Reference Line" located parallel to and 0.4-inch away from the zero line of reference; and (3) a scope with a multiple system of parallel lines, separated by tenths of an inch, each line representing 25 scaled feet.

For practical reasons, the errors were all reduced to percentage values in this section of the data, using only pips which were 50 or more scaled feet from the zero line of reference. Figures 1 to 3 are based on the composite records of five subjects. (It will be shown later that individual differences in random error are small.)

Figure 1 shows that, for the unscaled scope, the standard error was 11.09% of the space being estimated. Figure 2 shows that the use of
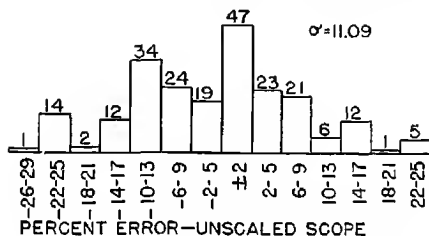
Fig. 1. Distribution of errors on the unscaled scope.
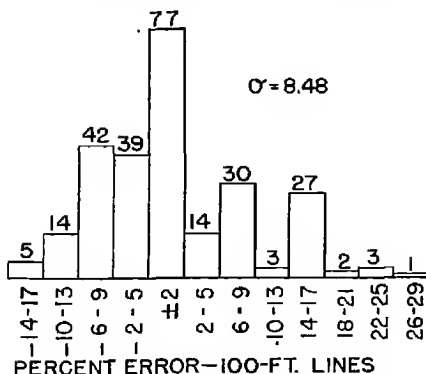
Fig. 2. Distribution of errors on the scope with 100-ft. reference lines.
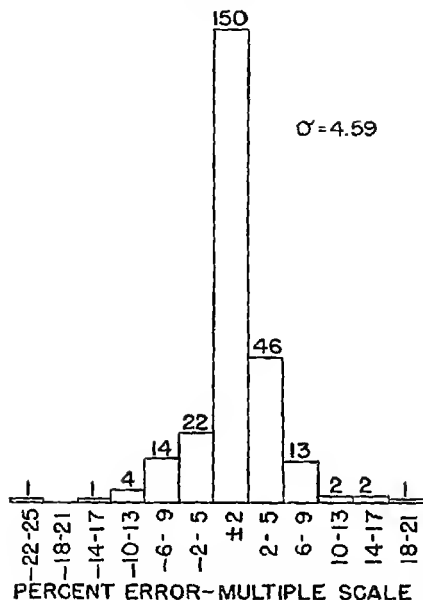
FIG. 3. Distribution of errors on the scope with multiple scaling.

side lines, 0.4-inch away from the zero line of reference, reduced the standard error to 8.48%. Figure 3 shows that with the use of a multiple system of lines, one tenth inch apart, the standard error is now reduced to 4.59%.

Now Garner (1) has shown that on PPI-type scopes, with scaling in the form of concentric rings, scaling of the degree of fineness in our multiple system produced confusion errors, decreased accuracy and promoted longer reaction times. We shall substantiate Garner's statement with respect to confusion errors, but we shall have to indicate, from evidence in Figures 1 to 3, that the smallest spread of *random error* was produced for the finer scaling. We found no statistically reliable difference in verbal reporting reaction time. This may be a difference

between human reactions on polar scaling, which Garner used, and rectangular scaling, which we used.

At this stage in the experiments we went into a more detailed gathering of data on the finely scaled scopes, to see whether or not the advantage of a smaller random error was not offset by the presence of systematic errors which could not be tolerated.

Absolute Amount of Random Error

Since we have ascertained that the more finely scaled scope yielded the smallest random error, in percentage figures, we shall now confine our measurements to the absolute values in this scaling situation (lines in tenths of an inch, representing 25 scaled feet of elevation, with 100-foot lines emphasized).

In Tables 1 and 2 the standard deviation of the error spread is computed omitting the confusion errors around the 100-foot scaling line, which are obviously not random. Mistaken numerical interpretations around the 25-foot scaling line cannot be distinguished easily from random errors, but we shall make an attempt, later, to show they exist by statistical analysis.

Individual differences, for untrained subjects on group experiments, with clear, uniform signals, are presented in Table 1. It appears safe to

Table 1

Random Error, Standard Deviation, Group Experiments, Individual Differences

In the following table the subjects are arranged in the order of best to worst, and all are untrained. The scaling consists of the multiple system with lines a tenth of an inch apart.

Subject	Stand. Dev. in Scaled Feet	Stand. Dev. in Scope Inches	Number of Readings	Subject	Stand. Dev. in Scaled Feet	Stand. Dev. in Scope Inches	Number of Readings
M. J. D.	2.8	.011	89	R. K. S.	4.9	.019	89
N. J. R.	4.1	.016	89	C. A. W.	5.2	.021	87
J. H. J.	4.5	.018	114	C. E. F.	5.5	.022	90
L. E. K.	4.5	.018	90	M. K. S.	5.6	.022	88
R. B. T.	4.5	.018	89	P. A. W.	5.8	.023	86
A. R. F.	4.6	.018	88	K. M.	6.0	.024	110
W. J. K.	4.6	.018	64	M. S. W.	6.1	.024	88
A. W. R.	4.7	.019	89	D. L. H.	6.5	.026	88
A. P. R.	4.8	.019	63	B. J. J.	7.0	.028	87
M. B. C.	4.9	.019	90				

Note: Confusion errors at the 25-foot minor scaling line cannot be accurately separated from random errors. The above standard errors include these, and are probably all too large. See Table 3 for an attempt at separation.

say, from these data, that average intelligent operators should be able to report elevation deflections to a standard error of a plus-or-minus 0.020 inch of scope distance, under such conditions. This represents an error in judging the elevation of a plane of five or six feet, presumably trivial. Trained subjects are much more nearly alike in error spread, and we have combined the runs in Table 2 to show the absolute error under six different experimental conditions.

Table 2

Distributions of Errors under Various Conditions, Elevation Reporting,
Multiple Scaling, All Subjects Combined

For a description of the character of each run, as designated by A, B, C, D, E, and F, see page 387 of the text.

Error, Scaled Feet	Character of Run						Location of Types of Errors
	(A)	(B)	(C)	(D)	(E)	(F)	
+110		1					Approximate band of confusion errors around the 100-foot major scaling line. Errors of overestimation.
+105		1		4			
+100	3	8	6	10			
+95		6		3			
+90				1			
+85				1			Approximate band of confusion errors around the 75-foot minor scaling line. Errors of overestimation.
+80		1					
+75				1			
+70							
+65							
+60							Approximate band of confusion errors around the 50-foot minor scaling line. Errors of overestimation.
+55						1	
+50		1					
+45						1	
+40		1					
+35				1	1	10	Approximate band of confusion errors around the 25-foot minor scaling line. Errors of oversetimation.
+30					7	19	
+25	3	1		6	13	29	
+20		4	3	9	38	97	
+15	2	7	2	7	106	135	Central band of random errors.
+10	56	67	76	58	159	174	
+5	268	370	375	365	262	193	
00	906	902	688	774	414	189	
−5	236	240	366	366	275	211	
−10	32	34	83	75	170	153	
−15	13	8	9	10	64	111	

Table 2 (Continued)

Error, Scaled Feet	Character of Run						Location of Types of Errors
	(A)	(B)	(C)	(D)	(E)	(F)	
−20	2	7	4	3	32	77	Approximate band of confusion errors
−25	3	7		3	19	27	around the 25-foot minor scaling
−30		4	1		3	13	line. Errors of underestimation.
−35			1		2	6	
−40		1				4	Approximate band of confusion errors
−45							around the 50-foot minor scaling
−50						1	line. Errors of underestimation.
−55		1			1		
−60							
−65							Approximate band of confusion errors
−70							around the 75-foot minor scaling
−75				1			line. Errors of underestimation.
−80							
−85							
−90	1						Approximate band of confusion errors
−95		2		3			around the 100-foot major scaling
−100	2	2		3			line. Errors of underestimation.
−105				1			
S.D. Feet	4.3	5.0	5.1	5.3	9.8	13.3	
S.D. Inches	.02	.02	.02	.02	.04	.05	

The six conditions in Table 2 are as follows:

Condition A. Five trained subjects. Individual experiments. Artificial scope with clear uniform signals. Rectangular presentation. Single-task elevation reporting.

Condition B. Nineteen untrained subjects. Group experiments before a large screen. Same problem materials as Condition A. Rectangular display. Single-task elevation reporting.

Condition C. Five trained subjects. Individual experiments. Artificial scope with clear uniform signals. Sector presentation. Single-task reporting.

Condition D. Nineteen untrained subjects. Group experiments before a large screen. Same problem materials as in Condition C. Sector display. Single-task reporting.

Condition E. Six trained subjects. Individual experiments. Simulator reproductions of field radar. Typical pip variations in contour,

size, brightness, shape, and hazy edges. Sector display. Single-task reporting.

Condition F. Six trained subjects. Individual experiments. Simulator reproductions of field radar. Same problem materials as Condition E. Sector display. Double-task reporting, alternating elevation reports with range reports.

The standard deviation of error distributions appears at the base of each colunm in Table 2, expressed both in scaled feet and in inches of actual scope distance.

Conditions A, B, C, and D all involve artificial scope pictures with clear, uniform signals. The conclusion that an average operator should be able to interpret distances, under these conditions, to a standard error of a plus-or-minus 0.020 inch is again substantiated. If a radar scope could be designed with such clear and uniform pips, and using scaling of this degree of fineness, this gives the human expectancy.

Condition E, using reproductions of an actual radar scope, shows that the random error is about doubled, due to signals which vary in shape, size, intensity, haziness of edges, etc. In the artificial series the reports were ten seconds apart. In this simulator series the operator reported every tenth pip, with the scan-line crossing the scope once every second. Rate of reporting was approximately the same, therefore.

Condition F is just like Condition E, except that the operator had to keep his attention on two tasks in alternation, elevation reporting and range reporting. The increase in standard error, from 9.8 feet to 13.3 feet, represents the effect of giving an operator an additional task. It may be presumed that the more tasks the radar operator is required to do simultaneously the less accurate he will be on each. This conclusion may seem to be something like proving the obvious, but it must be remembered that there is a proposal to make one man do what was previously done by from 3 to 5 men on GCA radar installations. The need for one-man operation is urgent, and the present study is merely an attempt to show that multiple tasks must be accompanied by extreme work simplification, if we are to avoid intolerable reporting errors. One confusion error, of the amount shown in Table 2 at the 100-foot line, could wreck an air transport.

Figure 6 shows the fit of a normal curve of distribution to the actual error distribution for the data of Condition E, reproductions of actual radar scopes.

Confusion Errors

Scales, both linear and surface types, consisting of major lines with numerical values, and minor divisions which are supposed to assist in
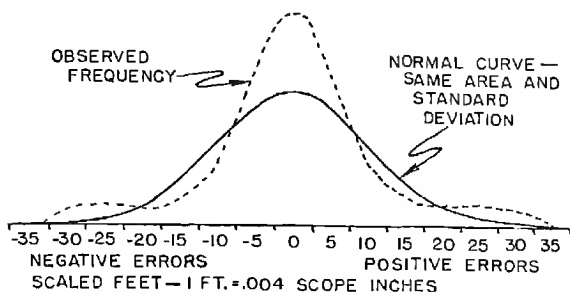
FIG. 4. Type of fit for normal curve when errors at 25 ft, the position of a minor scale division, have been included.

interpolation, are subject to mistaken interpretation of figures and errors in counting division points.

Table 2 shows a clear existence of mistaken interpretation at the 100-foot value.   This is verified by subjective reports, many times.   The 100-foot line is called a 200-foot line, or the line of zero reference is mistaken for a 100-foot side line.   There was no case of an error as great as 200 feet, but it was theoretically possible.

Also, at the 75-foot, the 50-foot, and the 25-foot distances there is an equal probability of assigning wrong numerical interpretations.   These
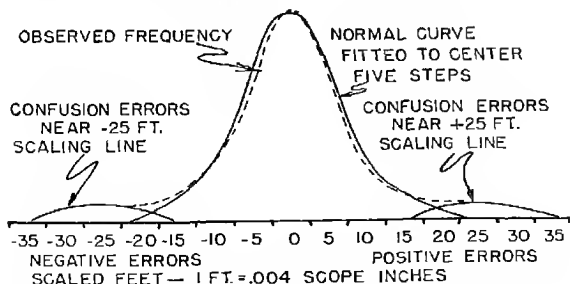
FIG. 5. Hypothetical improvement of normal curve fit when errors at the 25 foot scaling position have been excluded. Presented to explain the χ^2 improvement shown in Table 3.

are fairly clear at 50 feet and up. Unfortunately the confusion errors at the value of 25 feet overlap with the curve of random error. In fact, there is no way of separating confusion from random errors, at this position, but there may be a statistical way of showing facts which support the belief that they must be there.

Assuming that random error distributions should approach the curve of normal probability, an hypothesis which has considerable support, and that systematic errors will cause typical and expected distortions from normalcy, we may resort to the χ^2 test for these data. And in this use of the Fisher technique, it isn't just the bald fact that a misfit has occurred, but *where in the curve* the misfit is found, whether or not it is over the values which correspond to the minor or major scale points, that should
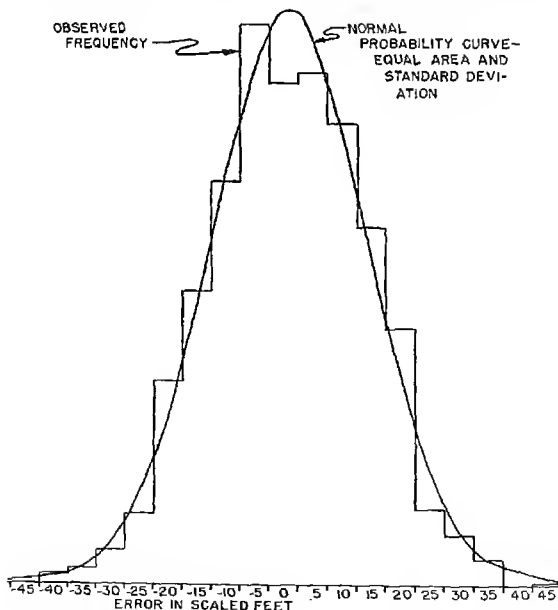
Fig. 6. Normal curve fit—random errors. Elevation reports in double task experiments.

prove of interest in spying out the presence of confusion errors mixed with random errors at the 25-foot distance.

Figure 4 shows the typical result we get when we try to fit a normal curve on our error distributions. The normal curve is plotted using the standard deviation of the distance from -35 to $+35$ feet, which includes confusion errors around 25 feet.

The χ^2 test always resulted in *too many errors* over the 25-foot position, and the discrepancy was *always positive* for every distribution beginning with Condition A through and including Condition E. This always produced the appearance of a leptokurtic hump at the center.

Table 3

Artificial Scope Runs
χ^2 Tests of Curve Fit for a Normal Distribution of Error,
Central Band of Random Error

In the following table the errors from -15 feet to $+15$ feet are hypothetically considered as being the band for pure random error (see Table 2), and that this region should fit a normal probability curve. The fit to the central band is tried two ways: with the supposed confusion errors included, and with the confusion errors excluded, i.e., by computing the standard deviation only on the central band.

Condition	Curve Area of Central Band	Stand. Dev. -35 to $+35$ feet	χ^2 Fit Central Band	Stand. Dev. -15 to $+15$ feet	χ^2 Fit Central Band	Number of Readings, Central Band
(A)	98.5%	4.31	167.25	3.40	80.19	1499
(B)	98.8%	5.10	24.28	4.65	5.89	1588
(C)	97.7%	5.01	225.97	3.70	48.67	1613
(D)	97.7%	5.30	97.01	4.40	9.92	1638
(E)	81.8%	9.80	39.79	8.50	23.27	1280
(F)	—	13.33	1.78	—	—	—

Note: The central band for Condition F was from -35 to $+35$, and was divided into five step intervals. No attempt was made to improve the fit because this was already as good as could be obtained. The area of this central band was 99.5% of the total distribution.

Figure 5 shows our hypothesis of what would happen if we determined the standard deviation by the central band of random error, only, and deliberately assumed that the excess of readings over the 25-foot point is due to confusion errors, not random errors.

Therefore, we adjusted the standard deviation value to fit the central band of error values, from -15 to $+15$ feet, and applied the χ^2 test again. The differences between the two assumptions are shown in final χ^2 answers in Table 3. Without exception, for the artificial scope series A to D inclusive, a χ^2 fit for 98% of the readings was greatly improved.

The astonishing thing was the discovery that the distribution for Condition F, double task reporting, was already almost a perfect normal curve, and could not be improved by any assumptions of systematic distortion.

We are inclined to believe, therefore, that the approximate bands for the regions of confusion errors in Table 2 are essentially correct. This means that, in reducing the random error by more finely divided scales, we have introduced an intolerable numerical confusion error, extremely dangerous for the practical navigation of aircraft by ground control radar. Therefore, no recommendation is made to use such a scale. More simplified methods of signal tracking must be designed, especially for one-man operation.

Persistence Errors

A rather broad definition of a persistence error may be: It is the tendency of an operator to bias a present report because of the mental persistence of a previous report.

We uncovered the existence of this possibility through two subjects whose data are plotted in Table 4. The first evidence was a sort of verbal stereotyping occurring when operators had to attend to two things alternately. Table 4 is drawn from the double-task experiments of Condition F.

Table 4
Distributions of Errors
Range Reports
Reproductions of GCA Field Radar Scope

Error in Scaled Miles	Initials of Subjects						Total	
	R. C.	D. M.	D. M. S.	J. H. F.	W. A. S.	C. B.		
+1.0					1		1	Band of
+.9					2		2	persistence
+.8								errors
+.7								
+.6					1		1	
+.5				1	1		2	
+.4								Band of ran-
+.3								dom and con-
+.2								fusion errors
+.1	13	29	17	31	32	17	129	
0.0	99	127	118	127	123	106	700	
—.1	84	50	68	46	49	70	367	
—.2	14		5	2		13	34	
—.3						1	1	

An operator would be reporting consecutive range values, "six-point-two, six-point-one, six-point-zero," and when he passed into the five-mile zone he went on, "six-point-nine, six-point-eight," and then suddenly remarked, "Oh, I meant five-point-eight." This is essentially the situation for Table 4.

This led us to wonder if something similar to this might not have been happening, to susceptible subjects, in the previous elevation serial reporting. Therefore, we computed the algebraic mean of errors following

Table 5

Trend of Algebraic Mean Error in Relation to Previous Report

Elevation Scale

A plus sign means that the subject tended to veer his reports in the direction of the preceding report. A minus sign means that the subject tended to bias away from the preceding report. The calculation is the difference in means where the preceding report was higher as compared with readings where the previous report was lower. Figures in italics are better than the one per cent level of significance. Differences are in scaled feet.

1. Group Experiments, Artificial Scope					
Subject	Difference	Subject	Difference	Subject	Difference
M. B. C.	+*2.38*	W. J. K.	+.32	N. J. R.	+.79
A. H. F.	−.24	L. E. K.	+.18	R. K. S.	+*2.05*
C. E. F.	+.45	K. M.	+1.56	R. B. T.	+*1.91*
D. L. H.	+.88	A. W. R.	−.61	C. A. W.	−1.37
B. J. J.	+1.18	A. P. R.	−.28	M. S. W.	+.21
J. H. J.	+1.07				
2. Individual Experiments, Artificial Scope					
L. A. A.	+*1.05*	F. P. H.	−.31	D. M. S.	+.84
W. A. S.	+1.02	R. J. R.	+*1.01*		
3. Simulator Reproductions of Field Radar					
C. R. B.	+*5.30*	J. H. F.	+3.07	W. A. S.	+4.60
R. C.	−2.24	D. M. S.	+.89		

larger previous values, and subtracted this from the algebraic means of reports following smaller previous reports. This difference is susceptible to a calculation for *reliability of differences of means.* Table 5 shows the results of this survey. Although only six out of twenty-six subjects showed a significance of difference better than the one per cent level, the general preponderance of plus values (20 out of 26) may carry some weight.

Granting that some subjects are susceptible to this effect, the size of

the error trend is actually too small to be of any serious consequence for the practical control of aircraft. A biasing effect of two feet, or even five feet, would not be intolerable. On range reporting it is conceivable that a mistake of one mile might be serious.

Summary

1. The use of finer scaling, with minor scale division to tenths of an inch viewed at sixteen inches, reduces random errors to a standard deviation of a plus-or-minus 0.020 inches of scope distances, for clear uniform pips, and 0.040 inches of scope distance for reproductions of actual radar pips.

2. The introduction of this finer scaling produces a proneness for confusion errors, defined as misinterpretation of the numerical values of scale positions. These errors may reach such a size as to endanger the navigation of an aircraft being guided by such operating reports.

3. Requiring an operator to alternate between two tasks in rapid succession has the effect of increasing the size of the random error, in our situation, about 30%.

4. Some subjects have a tendency to bias each report in a series by the mental persistence of the previous report. Only a minority of subjects do this consistently, and the amount is relatively small for practical significance.

5. Fine scaling, for one or more variables, is not recommended on the basis of present data for radar scopes.

References

1. Garner, W. R. Some Perceptual Problems in the Use of VG Remote PPI, Report of Research under Contract with the Office of Research and Inventions, U S Navy, 166-I-2, 15 September 1946. Restricted. The Johns Hopkins Psychological Laboratory. P. 34.

2. Ford, A. and Getz, M. H. Types of Errors in the Reading of GCA Scaled Scopes, Technical Report No. 4, Contract W28-099-ac-130, Watson Laboratories, Air Materiel Command, USAF, 31 August 1948. Restricted.

Some Design Factors in Making Settings on a Linear Scale *

William Leroy Jenkins and Minna B. Connor

Lehigh University

In setting a pointer on a linear scale by means of a control knob, is there an optimal ratio between pointer movement and knob turn? Is there an optimal knob diameter? When is a crank handle better than a knob? What is the effect of backlash in the system? No previous systematic investigation of such design factors seems to have been made.

The present study deals with a situation in which the operator is required to *match* a designated position on the scale with his pointer, rather than to set it to a specified numerical value. This limited phase of the problem was chosen because it permits data to be gathered rapidly and allows the accuracy of the setting to be objectively checked.

The primary criterion employed is the *time* consumed in making a setting, since time is comparable from subject to subject, and from condition to condition. In many of the experiments, action potentials from the active forearm were also picked up and measured. However, action potentials cannot be compared from subject to subject, since it is not known that the efficiency of the pick-up is the same in all subjects. For any given subject they do provide at least a rough indication of the relative amount of muscular work involved under different conditions.

Apparatus

Figure 1 is an operational diagram of the essential mechanical features of the apparatus. Rotation of the control knob turns the lower set of cone pulleys which drives the upper set of cone pulleys through a belt. Different ratios are obtained by shifting the belt. When the clutch is engaged, rotation of the upper shaft turns the drum and thus moves the pointer. When the clutch is disengaged, movement of the knob does not affect the pointer.

The linear scale consists of a black bakelite bar with vertical inserts of lucite .032″ wide at distances of 3, 9, 15, 21, 27, 33, 40, 56, 72, and 88 sixteenths of an inch symmetrically from the center. Behind each insert is a tiny flashlight bulb.

Through the center of the linear scale runs a thin metal strip which is used in checking the accuracy of setting. The pointer can be tipped to come in contact with the scale. If the pointer is entirely within the limits of a lucite insert, it will not touch the metal strip. If it is off the insert either way, it will come in contact with the metal strip and cause a red pilot lamp to light. The limit of error-tolerance is thus established by the width of the pointer.

The mechanical system is without backlash and is so adjusted that the pointer remains exactly where it was set after the clutch is released. To maintain these conditions, the belts must be quite tight; so that there is noticeable resistance at extremely coarse ratios. With the mechanical advantage of ratios in the medium and finer ranges, however, the operation requires very little effort.
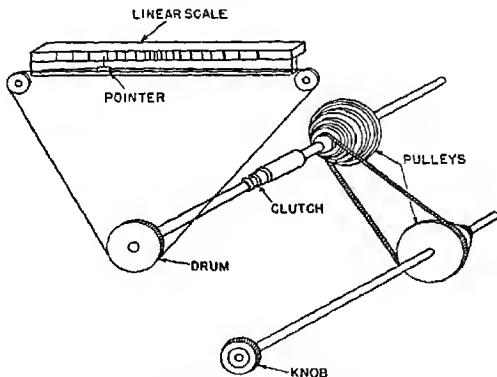
FIG. 1.   Mechanical features—operational diagram.

For measuring time, two chronoscopes are used; so that time for travel to approximate location and time for making the final adjustment can be separately determined. Similarly, two condensers are used to accumulate amplified action potentials during the travel and adjust phases. (Details of the electrical circuits and the four-stage amplifier will be found in the Technical Summary Report of the project.)[1]

[1] Jenkins, W. L. and Connor, M. B. Optimal Factors for Making a Setting on a Linear Scale, Technical Report No. 3, Contract W28-099-ac-130, Watson Laboratories, Air Material Command, USAF, 30 June 1948. Restricted.

Procedure

The procedure was essentially the same for all experiments. During a typical two-hour experimental session six or seven runs can be completed. Each run consists of a series of 20 settings, involving all 20 of the lucite inserts in a scrambled order. The procedure for a single setting is as follows:

1. After giving a preliminary warning signal, the experimenter closes a switch which simultaneously: (a) lights a pre-selected lucite insert; (b) starts both chronoscopes; (c) begins the accumulation of amplified action potentials in the first condenser.

2. As soon as he sees an insert light up, the subject starts turning the knob to bring the pointer from the center of the scale to the designated position. When the pointer comes within one tenth of an inch of the lighted insert, a contact is automatically closed which simultaneously: (a) stops one chronoscope; (b) switches the accumulation of action potentials from the first to the second condenser. Thus the first chronoscope and the first condenser provide measurement of the TRAVEL time and potential.

3. When the subject has completed his setting, he pushes the clutch release with his non-operating hand. This action simultaneously: (a) stops the second chronoscope; (b) cuts the second condenser out of the circuit. Thus the second chronoscope and the second condenser provide the ADJUST measurements.

4. The experimenter checks the accuracy of the subject's setting by tilting the pointer against the scale. (Errors occur so rarely that the very occasional "red light" reading is simply discarded.) The experimenter records the readings of both chronoscopes, and discharges each condenser separately into a sensitive ballistic galvanometer. The apparatus can then be reset for another trial.

Method of Analyzing Data

The raw data are in the form of time readings in tenth-seconds and action potential readings in arbitrary meter-scale units, for the travel and for the adjust phases of each setting. The adjust readings cause no difficulty because they can be averaged directly. However, travel readings vary according to the distance of the insert from the center. Hence, travel readings are first plotted against distance traveled and a straight line fitted. (The slope of this line is actually the travel rate, and the y-intercept an estimate of the starting time or potential.) Then the mean travel time (or potential) is scaled off for two standard distances: 10 sixteenths and 50 sixteenths of an inch. (The former is probably more representative of the usual amount of movement required in making

discrete adjustments.) Mean total time (or potential) = mean travel + mean adjust.

Subjects

Two former Navy radar operators (DMS and HWQ) were used in all of the experiments. Two other young men (JDS and RFM) with no such prior experience were available only for certain parts of the study. These four subjects are right-handed. The young woman (JKD) used in the study is naturally left-handed but was required to make settings with her right hand. She also had had no particular mechanical background.

Table 1

Influence of Ratio on Time and Potential

Standard Conditions

| | Mean Total Time | | | | | | | |
| | 10 Sixteenths Travel | | | | 50 Sixteenths Travel | | | |
Ratio	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
.220	25.2*	29.0*	24.0*	—	75.6*	66.6*	53.6*	—
.454	17.5	24.1*	23.1*	35.1*	30.5*	42.9*	37.9*	72.3*
.766	18.0	22.6*	22.4*	32.2	31.2*	35.4*	32.4*	52.7*
1.18	*16.5*	*18.6*	*19.4*	*30.3*	*24.3*	*22.7*	*25.8*	*44.8*
2.42	19.1*	21.6*	22.0*	29.1	27.1*	26.0*	25.6	38.7
4.08	19.2*	20.2	23.9*	35.4	23.6	24.6	27.9	42.2
6.28	19.5*	23.1*	26.7*	37.3*	23.5	27.5*	30.7*	43.3
9.70	23.8*	25.3*	28.1*	37.3*	26.6	28.9*	32.5*	42.1
16.3	32.8*	33.3*	37.2*	47.4*	34.4*	36.5*	42.4*	52.2*
33.6	54.3*	—	65.8*	—	57.9	—	73.0	—

| | Mean Total Potential | | | | | | | |
| | 10 Sixteenths Travel | | | | 50 Sixteenths Travel | | | |
Ratio	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
.220	24.3*	29.9*	26.9*	—	71.1*	78.7*	57.3*	—
.454	16.8*	20.8	19.5	27.3*	41.6*	46.8*	36.7*	64.5*
.766	15.3	19.5	19.0	22.1	28.5*	35.1*	29.4	42.9*
1.18	*14.4*	*19.7*	*20.5*	*20.3*	*23.2*	*23.1*	*23.3*	*36.3*
2.42	17.1*	16.4	21.2	17.5	25.1	22.6	26.8	26.3
4.08	16.5*	18.4	20.5	20.5	21.3	20.8	24.9	26.6
6.28	18.1*	16.4	21.9	25.8*	22.1	21.8	27.5	30.6
9.70	19.7*	18.4	22.6*	26.6	23.3	22.0	27.0	30.2
16.3	24.9*	23.4*	29.5*	33.4	26.9*	25.0	34.3*	37.8
33.6	25.4*	—	38.3*	—	28.1	—	43.9	—

* Significantly different from ratio 1.18.

Standard Conditions

The following conditions were standard in all experiments, unless specific exception is noted:

Linear scale—At eye level and normal reading distance.
Control knob—At waist level of seated subject; right-hand operation; 2¾" diameter knob.
Error-tolerance—.007" (pointer width of .025")
Ratios—Expressed in inches of pointer movement for one complete turn of the knob.

Mean total time is expressed in tenth-seconds for 10 sixteenths or 50 sixteenths travel distance. Mean total potential is expressed in meter-scale readings which have no absolute significance but are comparable for different conditions in the same subject. Each mean is based on a minimum of 80 readings. In tables showing italicized values, an

Fig. 2. Influence of ratio—standard conditions.
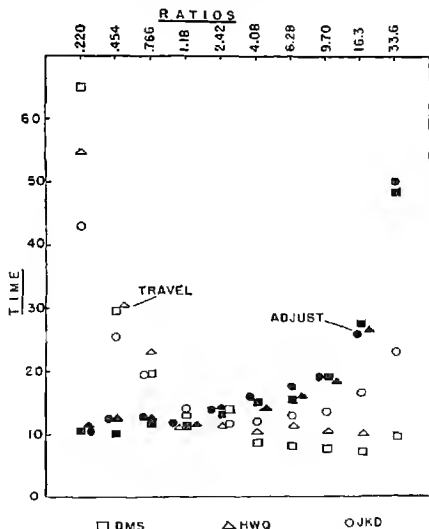
asterisk (*) indicates figures which differ significantly from the itialicized values, beyond the 1% level of confidence.

Results

Influence of Ratio. Is there an optimal ratio? Table 1 shows mean total time and mean total potential for ten ratios varying from .220 to 33.6 inches of pointer movement for one complete turn of the control knob. Although the subjects differ in their general levels, it is evident that the optimum is in the neighborhood of 1.18 in terms of both time and potential.

Figure 2 shows why the optimal ratio is in this region. For all subjects, travel time declines rapidly with increasing coarseness to about 1.18; thereafter coarser ratios do not speed up travel materially. In the opposite fashion, adjusting time declines with *decreasing* coarseness of ratio to about 1.18; thereafter finer ratios do not aid in making the final adjustment. A ratio about 1.18 combines rapidity of travel with speed of final adjustment.

For convenience in the remainder of this report we shall refer to 1.18 as "the optimal ratio." This should not be interpreted too literally. Actually there is an optimal *region* which holds good for all the subjects tested. Well-practiced subjects can use coarser ratios without undue loss, but the ratio designated as optimal has proved satisfactory for novice and expert alike.

Table 2

Stability of the Optimal Ratio

Standard conditions except that Feb. '47 figures were obtained with a 2″ diameter knob.

	Subject: DMS						Subject HWQ				
Ratio	Feb. '47	Apr. '47	May '47	Oct. '47	Mar. '48	Ratio	Feb. '47	Apr. '47	May '47	Oct. '47	Mar. '48
.220	28.1	—	20.6	25.2	—	.220	29.0	—	33.9	29.0	—
.454	20.3	—	22.4	17.5	—	.454	22.4	—	25.8	24.1	—
.766	18.3	—	20.7	18.0	23.2	.766	20.4	—	25.7	22.6	27.6
1.18	16.9	20.2	18.4	16.3	20.5	1.18	20.3	24.3	23.7	19.5	22.3
2.42	16.7	24.1	22.1	19.1	20.3	2.42	20.3	28.9	23.1	21.6	21.1
4.08	18.4	26.5	22.7	19.2	22.6	4.08	20.7	26.2	25.3	20.2	23.5
6.28	19.8	27.7	21.7	19.5	24.8	6.28	24.4	33.3	28.5	23.1	25.5
9.70	21.8	29.0	—	23.8	27.2	9.70	25.7	34.6	—	25.3	31.4
16.3	28.9	—	—	32.8	—	16.3	30.8	—	—	33.3	—
33.6	51.5	—	—	54.3	—	33.6	50.6	—	—	—	—

Mean Total Time 10 Sixteenths Travel

An indication of the stability of the optimal ratio over a period of time is presented in Table 2, which shows data for two subjects gathered on five different occasions over a period of thirteen months. Although the level of performance fluctuates from time to time, the optimal ratio remains in the same region.

Table 3 shows that the optimal ratio holds good for both the dominant and non-dominant hand. (To obtain these figures, a left-hand and a right-hand knob were coupled with auxiliary belts, so that the pointer could be set with either hand.) Particularly interesting here are the data for subject JKD. Although naturally left-handed, JKD had by this time become well-practiced in right-handed operation of the apparatus. At unfavorably high ratios she was now able to make faster settings with her right hand. Around the optimal ratio, the two hands were equally good.

Table 3

Ratios in Right vs. Left-Hand Operation

Standard conditions except that identical right and left hand knobs were coupled by a belt so that either could be used.

| | Mean Total Time 10 Sixteenths Travel | | | | | |
| | DMS | | HWQ | | JKD | |
Ratio	Right	Left	Right	Left	Right	Left
.766	22.2	24.4	25.5	29.5	25.0*	24.9
1.18	*21.8*	*24.6*	*24.8*	*28.0*	*22.5*	*23.6*
2.42	21.0	24.3	24.7	26.1	24.7	22.4
4.08	23.7*	25.0	25.1	26.4	27.6*	30.3*
6.28	26.0*	29.3*	29.8*	31.5*	27.7*	33.7*
9.70	29.2*	38.0*	37.6*	38.6*	31.6*	36.4*

* Significantly different from ratio 1.18.

Influence of Knob Diameter. In a preliminary study on two subjects, fourteen knob diameters were tested with five different ratios. For clarity in presentation the fourteen diameters are grouped in five step intervals. Table 4 gives the mean total time for 10 sixteenths travel distance. Several points of interest appear: (1) Regardless of knob diameter, the optimal ratio remains in the neighborhood of 1.18. (2) It is apparently not possible to compensate for an unfavorable ratio by altering the size of the control knob. Notice that the fastest times for ratio 6.28 are longer than the slowest times for ratio 1.18. (3) With coarse ratios the larger knob diameters work better. (4) At the optimal ratio, knob diameter appears to make very little difference.

As a check on this last point, five knob diameters were studied at the optimal ratio, using four subjects. Table 5 shows the results for both time and potential. In terms of mean total time, only the half-inch diameter is clearly unfavorable for all subjects, and the one-inch diameter mildly so for two of them. In terms of action potential, the $2\frac{3}{4}''$ diameter is significantly superior to the smaller sizes, although not always to the 4'' diameter.

Table 4
Interaction of Knob Diameter and Ratio

Standard conditions except that series of knob diameter were combined with series of ratios as indicated.

	Mean Total Time 10 Sixteenths Travel Subject HWQ				
Knob Diameters	Ratio 1.18	Ratio 2.42	Ratio 4.08	Ratio 6.28	Ratio 9.70
$\frac{1}{2}$, $\frac{3}{4}$	29.2	—	—	46.4	—
1, $1\frac{1}{4}$, $1\frac{1}{2}$	24.1	26.8	26.8	35.1	40.2
$1\frac{3}{4}$, 2, $2\frac{1}{4}$	22.6	25.3	25.6	31.2	34.2
$2\frac{1}{2}$, $2\frac{3}{4}$, 3	23.0	27.0	25.7	31.6	33.0
$3\frac{1}{4}$, $3\frac{1}{2}$, 4	24.3	27.3	25.0	30.8	30.7
	Subject DMS				
Knob Diameters	Ratio 1.18	Ratio 2.42	Ratio 4.08	Ratio 6.28	Ratio 9.70
$\frac{1}{2}$, $\frac{3}{4}$	21.5	—	—	34.1	—
1, $1\frac{1}{4}$, $1\frac{1}{2}$	21.5	24.3	30.0	37.5	33.7
$1\frac{3}{4}$, 2, $2\frac{1}{4}$	22.5	22.2	26.6	34.5	28.3
$2\frac{1}{2}$, $2\frac{3}{4}$, 3	21.6	22.5	26.3	28.4	29.6
$3\frac{1}{4}$, $3\frac{1}{2}$, 4	23.2	22.4	25.7	29.9	27.2

Figure 3 shows travel time and adjusting time separately. The half-inch diameter yields longer times for both travel and adjusting in all subjects. Among the larger sizes there is little to choose. It appears that the critical motion is the twist of the forearm, not the movement of the finger tips. Practically speaking, as long as the optimal ratio is used, the exact knob diameter does not matter, unless it is too small or too large to be grasped conveniently. The standard $2\frac{3}{4}''$ size used in most of our experiments was adopted simply because most subjects expressed a preference for this size.

Influence of Crank Handle. Cranks are generally used in tracking operations. The question has been raised whether a crank is better than a knob for making discrete settings involving large amounts of travel.

Table 5

Influence of Knob Diameter at Optimal Ratio

Standard conditions except that series of knob diameters were combined with ratio of 1.18.

	Mean Total Time							
	10 Sixteenths Travel				50 Sixteenths Travel			
Diam.	DMS	HWQ	JKD	RFM	DMS	DWQ	JKD	RFM
½	25.3*	28.1*	26.3*	42.1*	35.3*	38.1*	35.5*	53.7*
1	23.1	23.0*	22.0	39.3*	31.5	29.4	29.6	46.1*
2	21.1	22.9*	23.0	35.2	30.3	29.3	29.4	44.0*
2¾	*21.9*	*20.8*	*22.1*	*34.5*	*28.7*	*28.0*	*27.3*	*38.9*
4	21.2	21.8	21.8	37.6	27.6	29.4	26.6	46.6*

	Mean Total Potential							
	10 Sixteenths Travel				50 Sixteenths Travel			
Diam.	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
½	31.4*	29.2*	38.0*	33.4*	44.6*	40.0*	50.8*	46.2*
1	30.9*	24.2	33.0*	27.1*	44.1*	35.0	44.2*	35.9*
2	26.0*	25.5*	27.6	22.3*	38.4*	36.3*	39.2*	32.3*
2¾	*23.4*	*22.6*	*26.0*	*18.5*	*33.0*	*33.8*	*30.8*	*23.5*
4	21.7	26.7*	24.9	13.6	31.7	37.5	35.2	19.6

* Significantly different from 2¾.
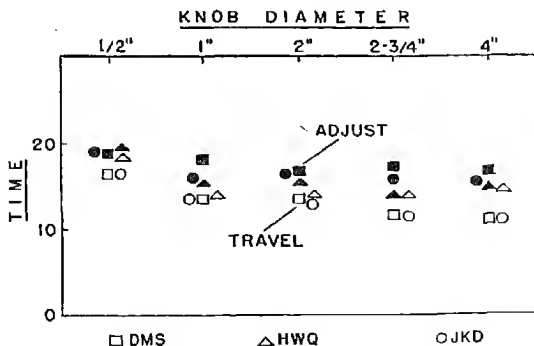
Fig. 3.   Influence of knob diameter—standard conditions.

To study this problem the $2\frac{3}{4}''$ knob was drilled so that a crank handle could be attached $\frac{1}{4}''$ from the periphery. Time measurements were taken at seven ratios under the following conditions: (1) Knob alone as a control; (2) crank attached and its use required; (3) crank attached but its use optional.

Table 6 shows mean total time for 50 sixteenths travel distance, which should give the crank the maximum advantage. Two interesting points appear: (1) Although the crank speeds up setting at ratios below 1.18, it does not enable these ratios to compete with the optimal ratio and the simple knob. (2) At the optimal ratio, the forced use of the crank is definitely deleterious and even its mere presence appears to hamper the best performance. Within the limitations of these experiments, at any rate, it appears that a crank handle serves no function whatever in making discrete settings on a linear scale.

Table 6

Comparison of Knob and Crank

Standard conditions except each mean based on a minimum of 40 readings. Crank simulated by attaching crank-handle to periphery of $2\frac{3}{4}''$ knob. In the table: KNOB means knob alone; CRANK means use of crank required; OPT means crank-handle present but use optional.

	Subject: DMS			Mean Total Time 50 Sixteenths Travel Subject: HWQ			Subject: JDS		
Ratio	KNOB	CRANK	OPT	KNOB	CRANK	OPT	KNOB	CRANK	OPT
.220	81.2	52.6	54.8	73.5	58.5	55.1	103.6	50.1	52.8
.454	52.7	35.6	36.7	48.0	42.5	39.9	64.6	40.1	38.7
.766	37.7	30.2	31.0	40.3	39.9	35.4	45.3	33.4	29.0
1.18	25.6	32.7	32.5	30.6	38.6	31.7	29.0	34.3	32.7
2.42	26.0	33.5	26.7	27.8	39.8	36.2	29.8	36.2	32.1
4.08	26.8	45.8	29.6	30.0	45.6	34.0	29.0	44.0	32.0
6.28	24.6	43.8	29.1	32.8	61.7	32.7	31.2	43.7	33.1

Influence of Backlash. Backlash is unavoidably present in some equipment. What is its influence on the speed of making settings? To study this question, the apparatus was modified by the addition of an arm moving between adjustable stops immediately beyond the subject's control knob, so that varying degrees of backlash could be introduced. In a preliminary series with two subjects, backlash was tested in 1° steps from 0° to 20° in the expectation that some particular amount of backlash might prove to be critical. Since this expectation was not realized, the figures have been grouped into seven step intervals. Table 7 shows mean total time for 10 sixteenths travel at ratios 1.18 and 6.28.

Surprisingly, backlash appears to have very little effect, even at the unfavorably coarse ratio of 6.28.

As a further check, backlash of 0°, 4°, 8°, 12°, and 16° was tested with three subjects using the optimal ratio. Results are given for mean total time and mean total potential in Table 8. Again it seems that no substantial effect of backlash can be found in either time or action potential. There is a slight upward trend with increasing backlash, but the statistically significant differences are scattered spottily and unconvincingly throughout the table. Figure 4 indicates that the slight upward trend comes from a minor lengthening of adjusting time, while travel time remains unaffected.

Table 7

Interaction of Backlash and Ratio

Standard conditions. Varying degrees of backlash introduced by means of an arm working between adjustable stops, immediately beyond subject's control knob.

| | Mean Total Time 10 Sixteenths Travel | | | |
| | Subject: DMS | | Subject: HWQ | |
Backlash in Degrees	Ratio 1.18	Ratio 6.28	Ratio 1.18	Ratio 6.28
0, 1, 2	23.1	27.8	24.4	29.2
3, 4, 5	23.2	30.1	24.9	28.1
6, 7, 8	23.8	32.5	25.8	28.7
9, 10, 11	25.4	33.0	26.4	30.1
12, 13, 14	25.1	32.7	26.4	32.2
15, 16, 17	26.1	32.5	26.2	30.7
18, 19, 20	26.5	33.3	26.6	29.7

We are reluctant to draw the sweeping conclusion that backlash is totally unimportant under all conditions. Perhaps with excessive friction or inertia, perhaps when far greater accuracy than .007″ is demanded, backlash may prove more disturbing than in the present experiments. Those are questions for further research to answer.

Influence of Error-Tolerance. How much does it slow up an operator to demand greater accuracy in setting? In our apparatus the error-tolerance could be altered simply by changing the width of the pointer in relation to the width of the lucite inserts. In a preliminary series, eleven pointer-widths were tested. Table 9 shows the results in terms of mean total time for 10 sixteenths travel distance. At the optimal ratio, only subject DMS shows a marked lengthening of time with decreasing tolerance; but at ratio 6.28 all three subjects show the same effect.

Table 8

Influence of Backlash at Optimal Ratio

Standard conditions. Varying degrees of backlash introduced by means of an arm working between adjustable steps immediately beyond subject's control knob.

Back-lash	Mean Total Time					
	10 Sixteenths Travel			50 Sixteenths Travel		
	DMS	HWQ	JKD	DMS	HWQ	JKD
None	*21.9*	*22.9*	*23.7*	*31.1*	*30.9*	*31.7*
4°	22.0	23.8	23.4	30.4	31.0	31.4
8°	23.4	25.5*	26.6	32.6	33.5	34.6
12°	24.2*	24.1	28.6*	34.2*	31.7	37.4*
16°	26.8	24.5	26.6	36.4*	33.3	34.6

Back-lash	Mean Total Potential					
	10 Sixteenths Travel			50 Sixteenths Travel		
	DMS	HWQ	JKD	DMS	HWQ	JKD
None	*25.7*	*23.9*	*32.9*	*38.9*	*36.7*	*43.7*
4°	28.0*	24.2	31.2	40.8	36.2	42.0
8°	26.4	26 6*	32.7	39.2	37 0	45.5
12°	26.7	25.3	33.9	39.5	37.3	46.3
16°	29.0*	26.5*	32.7	42.2	37.7	45.5

* Significantly different from None.
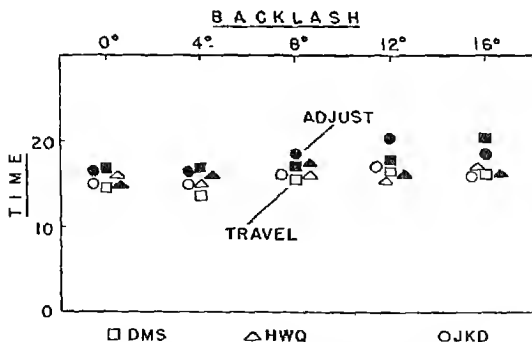
Fig. 4.—Influence of backlash—standard conditions.

Table 9

Interaction of Tolerance and Ratio

Standard conditions except that knob diameter is 2".

Error Tolerance	Mean Total Time 10 Sixteenths Travel					
	Subject: DMS		Subject: HWQ		Subject: JDS	
	Ratio 1.18	Ratio 6.28	Ratio 1.18	Ratio 6.28	Ratio 1.18	Ratio 6.28
.018", .016"	17.0	17.8	19.5	21.7	---	---
.013", .011"	16.5	21.0	20.7	23.2	22.1	27.9
.009", .008"	18.2	24.4	22.7	27.7	22.7	30.1
.007", .006"	24.2	28.6	22.7	30.0	24.2	30.0
.005"	26.5	37.1	24.1	29.5	29.9	33.4
.004"	30.0	52.1	25.2	33.2	32.7	39.9
.003"	35.3	50.2	29.0	39.2	33.9	40.5

A further study was made with four subjects, using five tolerances at the optimal ratio, measuring both time and potential. Table 10 gives the results. There is evidence of a moderate lengthening of time from .012" to .005"; then a sharp break at .003". From the reports of the subjects, it appears that .003" represents a breaking-point at which it

Table 10

Influence of Tolerance at Optimal Ratio

Standard conditions except that series of error-tolerances were tested at ratio of 1.18.

	Mean Total Time							
	10 Sixteenths Travel				50 Sixteenths Travel			
Toler.	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
.012"	15.8*	19.0*	16.6*	27.9*	22.8	25.4*	23.0	38.3*
.009	17.1	19.5*	18.3	31.4	23.9	26.3	24.7	40.2
.007	17.5	22.6	19.8	34.6	24.7	27.8	25.0	45.0
.005	20.7*	23.4	21.8*	38.1	27.9	31.0*	27.4	48.9
.003	27.7*	30.4*	25.9*	51.6*	33.3*	37.2*	32.3*	61.6*

	Mean Total Potential							
	10 Sixteenths Travel				50 Sixteenths Travel			
Toler.	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
.012"	14.1*	14.3*	21.4*	19.6	22.1*	22.7	30.6	30.0
.009	14.9	15.7	22.0*	19.5	23.2	22.9	31.6	29.9
.007	15.9	15.8	24.8	21.6	24.3	23.0	33.6	32.8
.005	17.2	19.6*	23.4	23.4	25.2	27.2*	31.8	34.6
003	21.5*	23.7*	27.6*	27.2*	29.0*	30.5*	36.4	37.6*

* Significantly different from .007.

becomes perceptually impossible to judge whether the pointer is accurately positioned. This is borne out by the fact that only at this level of tolerance did the subjects have an appreciable number of "red lights" (indicating that the clutch was released when the pointer was not within the confines of the lucite insert).

Figure 5 shows, as might be expected, that error-tolerance does not affect travel time. Adjusting time increases slowly as tolerance decreases, with a sharp upward break at .003″.
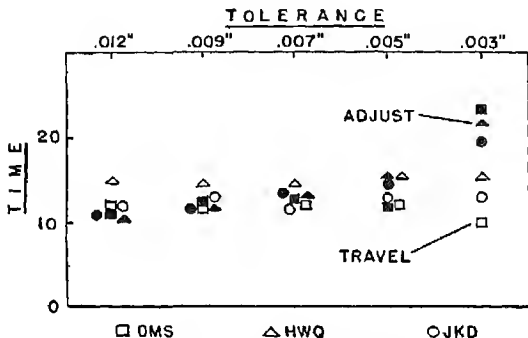
Fig. 5. Influence of tolerance—standard conditions.

It should be realized that .003″ represents a perceptual limit only under the conditions of this experiment; i.e., centering a pointer of appreciable thickness on a lighted insert. With ideal conditions, such as a fine hair line, it might be expected that the perceptual limit would be considerably lower.

Summary

In the foregoing experiments, the subject was required to move a pointer by means of a control knob and set it to a position on a linear scale indicated by a lighted insert. Time consumed in making the setting and the relative action potential developed in the active forearm were measured separately for travel to approximate location and for final adjustment. Systematic variations in ratio, knob diameter, backlash, etc., were introduced. Three to five subjects were used in the various parts of the study. The principal results follow:

1. The optimal ratio is one or two inches of pointer movement for one complete turn of the knob, for either the dominant or non-dominant hand. Finer ratios waste time and effort in traveling to the approximate location. Coarser ratios are clumsy for making the final adjustment. No other design factor investigated is as important as the optimal ratio.

2. Knob diameter is relatively unimportant, as long as the knob is large enough to be grasped conveniently. An unfavorably coarse ratio cannot be compensated for by altering the size of the control knob.

3. An unfavorably fine ratio cannot be compensated for by substituting a crank handle for the control knob. When the optimal ratio is employed, the addition of a crank handle to the knob does not aid and may be actually harmful, even when its use is optional.

4. Backlash, even in excessive amounts, has a relatively minor influence on either time or potential at the optimal ratio—under the conditions of this experiment. This may not be true under conditions of extreme friction and inertia, or when a tolerance much finer than .007″ is required.

5. Demanding greater accuracy of the subject by reducing the permitted error-tolerance increases time and potential only moderately, as long as the optimal ratio is employed. The final limit of accuracy in the present experiments appeared to be set by the perceptual difficulty of centering a pointer of appreciable thickness on a lighted insert, rather than by the limits of motor control.

Book Reviews

Lewin, Kurt (Edited by Gertrude Weiss Lewin). *Resolving social conflicts*. Selected Papers on Group Dynamics. New York: Harper and Brothers, 1948. xviii–230.

In his *Foreward* Gordon Allport writes such an excellent review of this book that the temptation to quote him liberally is too strong to be resisted. The thirteen papers, all previously published elsewhere are, he says, "so well-selected and so adroitly arranged that they provide an excellent introduction to Lewin's system of thought" (p. XIV). "The unifying theme is unmistakable: the group to which an individual belongs is the ground for his perceptions, his feelings, and his actions. Most psychologists are so preoccupied with the salient features of the individual's mental life that they are prone to forget it is the ground of the social group that gives to the individual his figured character. . . . This interdependence of the ground and the figured flow is inescapable, intimate, dynamic, but it is also elusive" (p. VII f.).

"Lewin's outstanding contribution is his demonstration that the interdependence of the individual and the group can be studied in better balance if we employ certain new concepts. Although the present volume contains primarily papers having a concrete, case-anchored character, still each shows with clarity how fruitful these new concepts are for understanding the phenomenon in question" (p. VIII.). Here, I think, we must be more cautious than Allport. We do not, it is true, quite share the objection sometimes made that Lewin's terminology is "metaphorical." All description consists in calling attention to similarities, and all terms are therefore metaphorical. What we should ask of the scientist proposing a new term is that he make clear the limits of generality involved. Is psychological or "life space" like geometric space in every respect? Lewin says it has all the qualities ascribed to space in non-quantified geometry (i.e., in topology). Presumably the life space has some but not all the characteristics of the more-familiar Euclidean space. Thus the term will for a long time have for us a strongly analogical coloration; it will suggest, that is, some properties which it does not have.

The merits of such a new way of describing facts must not, however, be overlooked. "Psychological or life space" suggests parallels which are actually confirmable hypotheses. The volume of significant research which has been set in motion by Lewin's array of terms is a tribute to their provisional utility. It is the reviewer's belief that they will

greatly illuminate also many of the social psychological problems in industry.

The more basic question comes when we consider the terms as explanatory concepts or constructs. Here fecundity in suggesting hypotheses is not an adequate criterion. Nor can we accept Allport's criterion of "understanding the phenomena in question." It is rare for concepts to seem unworkable in the concrete situation to which their own author seeks to apply them. A construct must prove itself in terms of stability in systematically varying conditions. In social psychology it may be years before the constructs can be tested in the requisite variety of critical situations.

Meanwhile, we do find here provocative interpretations of current problem situations. Part I deals chiefly with the problem of democratic re-education with particular reference to Germany. Part II deals with "Conflicts in Face-to-Face Groups." Part III, dealing chiefly with minority group problems, is somewhat more miscellaneous. The last chapter is significant because it reveals Lewin right up to the moment of his untimely death striving to see how, through action research, his hypotheses could be put to a genuinely experimental test. All persons interested in social engineering will find stimulation in this book.

Horace B. English

Ohio State University

Yoder, Dale, Paterson, Donald G., et al. *Local labor market research.* Minneapolis, Minnesota: University of Minnesota Press, 1948. Pp. xvii, 226. $3.50.

Early in 1939 officials of the city of St. Paul, Minnesota became aware of an apparent paradox. Although employment had been restored to proportions equal to those of the predepression period relief loads and expenditures continued at the high levels typical of the depression years. A Mayor's Committee on Unemployment studied the problem but found no satisfactory explanation. Finally the committee turned to social scientists at the University of Minnesota for help with the problem. In the early 30's the Employment Stabilization Research Institute of the University had made a series of significant studies of employment and unemployment and was thus uniquely equipped in 1940 to attack the immediate problem facing the city of St. Paul. The story of the research efforts of the ESRI during the years 1940–42 is reported in *Local Labor Market Research.*

The significance for psychologists of this account arises in part from the cooperative nature of the enterprise since the research staff included psychologists as well as economists, sociologists, and statisticians. Much of the methodology will interest applied psychologists in the fields of

opinion polling, counseling, and personnel administration. Finally, the findings, particularly of Project 3 constitute important new contributions to personnel psychology.

After a one-year pilot study it was apparent that the research program should include a comprehensive study of the labor marketing process. Five projects were selected for study.

Project 1 appraised available employment data particularly those of state and federal agencies and attempted to improve these labor market reports as a means of providing continuing indices of employment, hours, wage rates, and earnings. These measures were based on employer reports to various public and private agencies and covered only the employed.

Project 2 sought to provide detailed information on the numbers and types of labor supplies available and to serve as a check on the data obtained in the first project. In addition, special studies were undertaken to obtain information regarding priorities unemployment, civilian morale, nature and extent of vocational training, housing, shopping habits, transportation, and migration. The method was a continuous sampling survey using both a panel and randomly selected respondents in St. Paul. The result is an impressive demonstration of the use of sampling techniques in maintaining a continuing check on the dynamic elements of the labor market and in providing basic information on a wide range of community problems.

Project 3 concerned itself with some of the frictions in the labor market which interfere with the matching of men, women, and jobs. Psychological tests, interviews, and attitude surveys were among the tools used in studying the human factor in employment.

An attempt was made to relate available employment data to training opportunities available in the community. Data on school enrollments and the employment experiences of post-graduate youth were collected. The findings raised the question as to how well the public school system had fulfilled its responsibilities for vocational training.

Opinion polling methods were used to identify and measure attitudes and attitude changes among various occupational groups. Findings indicated that members of the labor market often held opinions at variance with the facts and this doubtless accounted for some of the labor force frictions. It was possible to get some idea of the job satisfaction of various occupational groups through this polling approach. Questions regarding public policy such as *"Do you think it is too easy for people to get on relief?"* got at attitudes which indirectly affect employment policies.

Of great interest to personnel psychologists is that portion of the study which compared the occupational classification assigned on the

basis of intensive clinical study to unemployed job seekers with those classifications routinely assigned by employment office interviewers. The results indicate that clinical study rather than a superficial appraisal based primarily on past job experience will identify a considerable number of persons whose potentialities for employment otherwise go undiscovered.

This intensive clinical study of almost four hundred unemployed persons yielded other useful information. For example, it was found that counseling letters may be useful in a large-scale counseling program where time for interviews is at a premium. Other analyses gave information on the dominant causes of unemployment.

A follow-up study of persons tested and studied clinically ten years previously indicated that occupational adjustment can be predicted with surprising accuracy. Re-tests on these same people gave amazingly high re-test correlations on pencil and paper tests being about .9. Correlations for performance tests were somewhat lower, being in the neighborhood of .6 to .7.

Project 4 was an attempt to tease out some of the complex interrelationships which influence the demand for labor. Analyses of economic data and opinion surveys were the methods used. The latter sought to secure and classify employers' opinions as to how and why they make decisions to offer employment. In a study of the printing industry the employees were also polled to ascertain any divergencies.

Project 5 was an analysis of relief administration policies and practices on the assumption that factors other than those of the labor market might be responsible for the St. Paul paradox of increasing employment without an accompanying decrease in relief rolls. Analyses of official reports of social work agencies provided one source of data. A major part of the study, however, was an intensive analysis of the characteristics of relief recipients. Finally, detailed study was made of fifteen relief clients for whom a great deal of information was available as a result of their participation in the occupational analysis work of Project 3.

The findings of Project 5, as a whole, indicated that the nature and conditions of relief administration were an important factor in the situation. It seemed clear that relief expenditures reflected much more than the current condition of local labor markets.

Both the conduct of this research program and the nature and form of publication were materially affected by the war. Changes in personnel and finally the withdrawal of foundation support because of war conditions brought the study to an end before it really was completed. Thus this book is more of a progress report emphasizing methodology than it is a definitive statement of the findings. The compilation and publication

of this report actually was undertaken by the Industrial Relations Center established at the University of Minnesota in 1945. It is the work of many authors and reflects some of the obvious limitations. Credit for a careful editing should go, however, to Herbert G. Heneman, Jr.

This is a unique and important contribution to labor market research and is a milestone marking the road which psychology is traveling toward cooperative research on meaningful problems. At a time when "action research" has become a fashionable term among social scientists the reviewer judges this report to be a significant demonstration of the application of psychological viewpoint and methodology to pressing social problems. This categorization as action research, be it noted, is expressed at the risk of embarrassing the directors of the enterprise who have long engaged in the study of real problems without benefit of a more esoteric terminology applied to their highly productive efforts.

Arthur H. Brayfield

University of California

Jucius, Michael J. *Personnel management.* Chicago: Richard D. Irwin, Inc., 1948. xii + 696 pp. $6.00.

Personnel management is defined as "the field of management which has to do with planning, organizing, and controlling the performance of various activities concerned with procuring, developing, maintaining, and utilizing a labor force such that the objectives and purposes for which the company is established are attained as effectively and economically as possible, and of labor itself are served to the highest possible degree."

Around this definition Jucius has written a college textbook designed to provide a "realistic study of the principles and practices of personnel management." The thirty chapters deal systematically with organizational problems, approaches, and techniques in selecting, training, renumerating, and motivating employees and in maintaining satisfactory labor-management relations.

The presentation is in the typical textbook fashion. It is well-organized and will lend itself to outlining in the student's notebook. The emphasis seems to be upon presenting a body of material to be studied under the guidance of a qualified instructor rather than upon providing a self-motivating treatise for the general reader. It differs from the standard textbook, however, in that supporting source materials are rarely given. Footnote references are infrequent and there are no suggested additional readings for separate chapters.

The chief merit of the text is its well-organized and systematic presentation of a wealth of information about personnel practices and principles. It is chuck-full of step-by-step procedures, examples of forms, and practical suggestions for approaching the common problems faced by a

personnel department. It emphasizes the importance of "getting the facts" and of careful follow-up and control after the appropriate steps have been taken. The final chapter stresses the need for a research point-of-view which would lead to continuous intensive study of all aspects of personnel management.

Since this review is written by a psychologist for psychologists it is pertinent to look for evidence of the impact of psychological findings upon personnel practices as described, even though the author is not writing a text on personnel psychology. In this respect the presentation is rather weak. Recognition is made of individual differences and of the importance of employee attitudes and feelings and, rather frequently, some rather cogent observations on human nature are reflected in common sense statements. There is little overt recognition, however, of the dynamic nature of interpersonal relationships, of the fundamental problem of democracy in industry, of the individual as a person rather than as an employee. The areas in which psychology has made specific contributions in industry are the most poorly presented, viz., interviewing, counseling, and testing. The influence of the social structure in company organization is not described, the Hawthorne studies being referred to merely as an example of research.

In summary, "Personnel Management" will serve as an excellent textbook in the field of business administration if supplemented by source materials, if livened up by a stimulating instructor, and if the students also take courses in personnel and industrial psychology.

Albert S. Thompson

Teachers College,
Columbia University

Lall, Sohan. *Mental measurement.* Allahabad: Allahabad Law Journal Press, 1948. Pp 88.

This little book presents results obtained from the administration of three tests to approximately 2000 Indian children in 58 government high schools. The children were 11+ years old and the tests were a group verbal intelligence test, an English language test, and an arithmetic test. No data are given on the construction of any of the tests; the author states simply that they were patterned after the Moray House tests of Godfrey Thomson. The tests in English and arithmetic were achievement examinations in these areas.

Distributions of test scores in the entire sample are presented and the method used for removing the skewness which appeared in all three distributions is defended. Perhaps the most interesting part of the monograph is the presentation of comparative test scores for four Indian castes, for children from different geographical regions, and for children

whose parents fell into various occupational groupings. Most of the differences are quite small but some are statistically significant. It seems likely, however, that the apparent significance was, in many cases, a resultant of the small and probably unrepresentative samples. How representative the samples were we have no way of judging.

On the whole the monograph is somewhat amateurish and reminds one of publications in this country of some 25 years ago. This is a pioneer job, however, done under considerable difficulties, and the author deserves a great deal of credit. The references in the book are to Thomson, Spearman, and Burt, under whom the author apparently had his training.

<div align="right">Henry E. Garrett</div>

Department of Psychology
Columbia University

Evans, Ralph M. *An introduction to color.* New York: John Wiley and Sons, 1948. Pp. x + 340. $6.00.

Any serious treatise on color is a major undertaking which necessitates the coordination of materials from physics, physiology, and psychology. This book was written with the avowed purpose of giving adequate treatment to materials from each of these three fields. Each phase is treated separately and then the three are interwoven near the end of the book. Consistent, understandable terminology is achieved by employing common speech meanings of words, with a minimum number of new words introduced and defined. Many pictures and graphs are employed to help the reader grasp the fundamentals. To a large degree the text is descriptive and non-mathematical. Although it is not assumed that the reader has more than an elementary knowledge of physics and psychology, no simplifying omissions of subject matter are made. The author, head of the Color Control Department of the Eastman Kodak Company, is attempting to give the reader the benefit of his twenty years practical experience in the field.

In this book, the author has been fairly successful in achieving his aims. The material is not a popular treatise, but a simplified technical discussion of highly complex and technical subject matter. Although not easy reading, persistent study of the material will be found rewarding. It is the only book known to the reviewer that attempts to give such a complete story of color. There is somewhat more emphasis upon physiological and physical than upon psychological aspects. Nevertheless, the psychologist will profit greatly by reading the book. Especially he will be able to correct many inaccurate notions obtained from elementary discussions.

One wonders why a discussion of geometric optical illusions are in-

cluded in this treatise on color. Furthermore, to include the ambiguous staircase as an illusion is erroneous. The book would be more complete if a thorough discussion of color experiences of partially (red-green) color blind persons were included. Another item that would improve the treatise is a more complete discussion of color in illumination, and lighting in relation to color in interior decoration.

Some of the more important sections deal with the use of colors in photography, art and display situations. In general, this book is well organized and clearly written. It will be useful both to those interested in the fundamental principles of color and to those working with color applications in practical situations.

Miles A. Tinker

University of Minnesota

New Books, Monographs, and Pamphlets

Guiding human misfits. Alexandra Adler. New York: Philosophical Library, 1948. Pp. 114. $2.75.

The psychology of development and personal adjustment. John E. Anderson. New York: Henry Holt and Co., 1949. Pp. 720. $3.25.

Fatigue and impairment in man. S. Howard Bartley and Eloise Chute. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 429. $5.50.

Psychological factors in education. Henry Beaumont and Freeman G. Macomber. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 318. $3.00.

Psychology of personnel in business and industry. Roger M. Bellows. New York: Prentice-Hall, Inc., 1949. Pp. 499. $4.50.

A summary of clerical tests. George K. Bennett and Ruth M. Cruikshank. New York: The Psychological Corporation, 1949. Pp. 122. $1.25.

Encyclopedia of criminology. Vernon C. Branham and Samuel B. Kutash, Editors. New York: Philosophical Library, 1949. Pp. 527. $12.00.

Psychological tests for retail store personnel. Dora F. Capwell. Pittsburgh: Research Bureau for Retail Training, University of Pittsburgh, 1949. Pp. 48. $1.00.

Reading manual and workbook. Homer L. J. Carter and Dorothy J. McGinnis. New York: Prentice-Hall, Inc., 1949. Pp. 120. $1.75.

The psychology of social classes. Richard Centers. Princeton: Princeton University Press, 1949. Pp. 256. $3.50.

Applied experimental psychology, the psychology of engineering design. Alphonse Chapanis, Wendell R. Garner, and Clifford T. Morgan. New York: John Wiley and Sons, Inc., 1949. Pp. 402. $4.50.

Introduction to the Szondi Test. Susan Deri. New York: Grune and Stratton, 1949. Pp. 354. $5.00.

Practical lessons in psychiatry. Joseph L. Fetterman. Springfield: Charles C. Thomas, Publisher, 1949. Pp. 342. $5.75.

The art of readable writing. Rudolf Flesch. New York: Harper and Brothers, 1949. $3.00.

Adolescence. C. M. Fleming. New York: International Universities Press, Inc., 1949. Pp. 261. $4.50.

The energetics of human behavior. G. L. Freeman. Ithaca: Cornell University Press, 1949. Pp. 350. $3.50.

Workbook manual for marriage and the family. Revised edition. John Harvey Furbay. New York: Appleton-Century-Crofts, Inc., 1949. Pp. 248. $2.00.

American social reform movements: their pattern since 1865. Thomas H. Greer. New York: Prentice-Hall, Inc., 1948. Pp. 313. $4.00.

Elmtown's youth. A. B. Hollingshead. New York: John Wiley and Sons, Inc., 1949. Pp. 480. $5.00.

Adolescent development. Elizabeth B. Hurlock. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 566. $4.50.

Learning to drive safely. A. R. Lauer. Minneapolis: Burgess Publishing Co., 1949. Pp. 145. $2.25.

Communications research 1948–1949. Paul F. Lazarsfeld and Frank Stanton. New York: Harper and Brothers, 1949. Pp. 332. $4.50.

Older people and the church. Paul B. Maves and J. Lennart Cedarleaf. Nashville: Abingdon-Cokesbury Press, 1949. Pp. 272. $2.50.

The effect of experience on nursing achievement. R. Louise McManus. New York: Bureau of Publications, Teachers College, Columbia University, 1949. Pp. 64. $2.10.

Psychiatry: its evolution and present status. William C. Menninger. Ithaca: Cornell University Press, 1949. Pp. 149. $2.00.

Genetics, medicine, and man. H. J. Muller, C. C. Little, and Laurence H. Snyder. Ithaca: Cornell Univeristy Press, 1949. Pp. 164. $2.25.

An introduction to clinical psychology. L. A. Pennington and I. A. Berg, Editors. New York: The Ronald Press Co., 1949. Pp. 600. $5.00.

Education through art. Herbert Read. New York: Pantheon Books Inc., 1949. Pp. 320. $5.50.

Psychodiagnosis. Saul Rosenzweig. New York: Grune and Stratton, 1949. Pp. 380. $5.00.

The clinical application of psychological tests. Roy Schafer. New York: International Universities Press, Inc., 1948. Pp. 346. $6.75.

Problems of early infancy. Milton J. E. Senn, Editor. Second Conference of Josiah Macy, Jr. Foundation. New York: Josiah Macy, Jr. Foundation, 1948. Pp. 120. $1.00.

Individual behavior. Donald Snygg and Arthur W. Combs. New York: Harper and Brothers, 1949. Pp. 386. $3.50.

Learning theory in school situations. Esther J. Swenson, G. Lester Anderson and Chalmers L. Stacey. Minneapolis: University of Minnesota Press, 1949. Pp. 103. $1.50.

Thematic apperception test. Charles E. Thompson. Cambridge: Harvard Univeristy Press, 1949. Manual $.50, Test $5.00.

Man's quest for significance. Lewis Way. New York: The Macmillan Co., 1949. Pp. 211. $3.50.

The inner world of man. Frances G. Wickes. New York: Henry Holt and Co., 1949. Pp. 320. $5.00.

Trends in student personnel work. E. G. Williamson, Editor. Minneapolis: University of Minnesota Press, 1949. Pp. 417. $5.00.

Jobs and the man: a guide in understanding and dealing with workers. Luther E. Woodward and Thomas A. C. Rennie. Springfield: Charles C. Thomas, Publisher, 1946. Pp. 125. $2.00.

Occupational outlook handbook. Bureau of Labor Statistics, Bulletin 1949, No. 940. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1949. $1.75.

Guidance handbook for elementary schools. Office of Los Angeles County Superintendent of Schools. Hollywood: California Test Bureau, 1948. Pp. 158. $2.40.

Guidance handbook for secondary schools. Office of Los Angeles County Superintendent of Schools. Hollywood: California Test Bureau, 1948. Pp. 246. $3.00.

An Objective Analysis of Morale

William James Giese

William James Giese, Ph.D. and Associates, Chicago 3, Illinois

and

H. W. Ruter

Aldens, Inc., Chicago 7, Illinois

The profit and loss statement of a business is affected by that elusive thing called morale. Since most successful executives have recognized this, many companies have attempted to get a workable measure of the status of morale among their employees. The most successful of these attempts has been the morale survey through the use of a correctly designed questionnaire to measure the attitudes of the employees toward their supervision, working conditions, wage rates, chances for advancement, and similar important attitude areas. Although the results of such morale surveys are usually interesting to management (sometimes the results are even startling), the questionnaire method is cumbersome, costly, and slow. Also, because of its nature, the morale survey can be given successfully only at intervals of about once a year. This limitation, in addition to the costliness, often prevents the detection of an undesirable trend in morale when corrective action is easiest and most effective.

In addition to these very serious limitations, the morale survey becomes a row of question marks when a cost and savings analysis of it is made. The only answer to the question of costs is a general agreement among executives that poor morale costs the company money. But just how much money has not been determined, for there are no accounts in which poor morale shows up as an identifiable and cost accountable loss. As business moves into a more competitive era this question of costs and savings presents itself with an ever increasing urgency. If increasing the morale of the employees costs money, how much will be saved for every dollar spent? Is there a straight line relationship; that is, are the savings per dollar spent the same regardless of the amount spent? Where is the break-even point? At what point does the law of diminishing returns begin to operate?

After considering the above facts and questions along with their implications, the personnel manager [1] began to build an index based on objective records. After a number of conferences we organized a program of research that would give us some of the answers and tie the results directly to employee morale.

Purpose

Our primary purpose was to analyze the relationship of the objective records of departmental performances to morale as measured by the questionnaire method. Once these relationships are known, it is merely a mathematical problem to determine the best relative mathematical weights for each of the factors for the purpose of predicting morale.

In order to make certain that our basic records held promise of providing fundamental data for the prediction of morale, we first examined these factors as well as the morale score itself. Since the morale questionnaires were not scored but had only percentages of responses for each part of an item, it was necessary to make up a scoring system for the morale questionnaire.[2]

The first step was to learn if on the basis of our scoring there was an adequate range of difference between the departments in morale. Similar analyses were made for six factors [3] on which objective data were available in the personnel department and which appeared feasible for study to determine their prediction value in indicating departmental standings.

Our final purpose was to set up a simple method for the determination of departmental morale based upon factors each of which are cost accountable.

[1] H. W. Ruter, Personnel Manager of Aldens, Inc., 511 S. Paulina Street, Chicago 7, Illinois. Aldens is a national mail order company which retails by mail and through department stores all types of clothing, housewares, furniture, etc.

[2] The morale questionnaire was scored by giving the most unfavorable response to an item a value of 1, the second most unfavorable response a value of 2, and so on up to 6 for the most favorable response. Seventeen of the eighteen items in the questionnaire were amenable to this scoring system. The morale score was merely the total points for the seventeen items.

[3] The factors considered were: (a) Seasonal accumulated Departmental Production Efficiency; (b) Seasonal accumulated Departmental Error Efficiency, covering errors not affecting customers, i.e. errors in company handling which may delay the completion of the sale but are not otherwise noticeable to the customer; (c) Seasonal accumulated Departmental Error Efficiency, covering errors affecting customer such as items charged and omitted, wrong merchandise, size, color, etc. This type of error is most costly because of its injurious effect on customer relations, merchandise loss, etc.; (d) Annual labor turnover rate; (e) Seasonal accumulated tardiness percentage; and (f) Seasonal accumulated absence percentage.

The Results

The morale score gave an adequate separation between departments. In Figure 1, each square represents a department, and the number below the square is the average (arithmetical mean) morale score of the employees in the department.
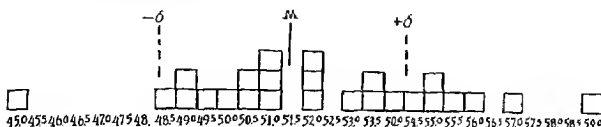
FIG. 1. Distribution of departmental morale scores.

In productive efficiency the mean of the departments varied from 80% to as high as 113%. This means that there are large enough differences between departments in their productive efficiency averages to use this factor in the correlation analysis. Figure 2 illustrates departmental differences in average productive efficiency.
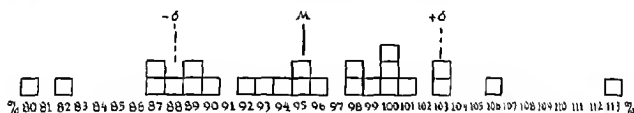
FIG. 2. Distribution of departmental averages on per cent productive efficiency.

For the remaining 5 factors the departmental differences are summarized in Table 1.

Table 1

Means and Standard Deviations of the Departmental Means for the Six Objective
Factors used in the Analysis of Morale

Factor	Mean	S.D.
Per Cent Production Efficiency	95.64	7.43
Per Cent Error Efficiency Not Affecting Customers	78.41	20.85
Per Cent Error Efficiency Affecting Customers	61.60	17.00
Per Cent Turnover	222.70	80.02
Per Cent Late	9.67	2.87
Per Cent Absent	7.53	1.94

The present error efficiency **NOT** affecting customers has a cluster of departments averaging between 84% and 94% (14 out of 22) but due to the fact that the entire range is from 20% to 100% we kept this factor for correlational purposes.

The per cent error efficiency affecting customers shows a wide range of departmental performances and an even spread of differences throughout the range. Such a condition makes this factor a promising partial for the prediction of morale. Also, when the cost analysis is made of such errors, the potential savings due to reduction of these errors will in all probability be many times those possible with errors **NOT** affecting customers. The reason for this is that errors affecting customers are not only more costly but that there are only a few departments with a high efficiency standing.

The highest departmental labor turnover was 350% and the lowest 40%, almost a 9 to 1 ratio. The costliness of turnover can easily amount to a six or seven figure annual loss since the minimum loss per employee termination is $100.00. In addition, this factor for the purpose of predicting the morale standing of a department should be of great importance because of the large and even spread of departmental standings.

There are wide departmental differences in the per cent of the employees who are late. Here the range is from 3.5% to 14.5%; a ratio of well over 3 to 1. This factor also has an adequate spread and evenness throughout the spread for correlation analysis.

Since a department must carry more personnel to meet the work load demands if absenteeism is high, departmental variations in per cent absent were analyzed. The lowest was 3½% and the highest was 11%; a ratio of almost 3 to 1. The department which *averaged* 11% absent for the year had to carry at least 11% more employees to meet adequately the work demands placed upon it.

Table 2 shows the results of correlating each of the six factors with the morale score and with each other.

The two factors with the highest relationship to morale are per cent turnover and per cent absent. Both relate to morale with a fair *negative* relationship. When morale is low, absenteeism and turnover tend to be high.

Error efficiency affecting customers has a higher relationship with morale than does error efficiency **NOT** affecting customers. Most interesting is the low *negative* relationship between these two factors. That is, those departments with a high efficiency in errors affecting customers have a slight tendency to be low in errors **NOT** affecting customers. This difference could be accounted for on the basis of departmental emphasis on the importance between the two types of errors. There is a

slight *negative* relationship between morale and per cent late in a department. Since *outside factors such as number of transfers made on public transportation, distance from place of work, weather, etc.* probably have little or no relationship with morale but a fairly high one with lateness, this low relationship between morale and per cent late is to be expected. There is a slight relationship in the positive direction between morale and production efficiency, but it is much less important to morale than per cent turnover or per cent absent. This means that high morale is only

Table 2

Pearson Product-Moment Coefficients of Correlation between Objective Records of Departmental Performance and the Morale Score *

	0. Morale Score	1. % Productive Efficiency	2. % Error Efficiency NOT Affecting Customers	3. % Error Efficiency Affecting Customers	4. % Turnover	5. % Late
1. Per Cent (%) Productive Efficiency	+.19					
2. % Error Efficiency NOT Affecting Customers	+.15	−.50				
3. % Error Efficiency Affecting Customers	+.27	+.37	−.24			
4. % Turnover	−.42	−.18	+.05	−.25		
5. % Late	−.20	−.18	+.30	−.28	+.33	
6. % Absent	−.47	−.15	−.18	−.07	−.15	−.03

* The above correlations were computed from the averages of 25 departments. Therefore the scattergram was composed of 25 points. The total number of employees represented by the 25 points was 3000. The number of employees in the departments ranges from 14 to 405 with a mean of 120 and a sigma of 90.5. The correlation between the morale score and the number of employees in the department was −.07.

slightly related to per cent productive efficiency which makes it is entirely possible for a department to have a high standing from the standpoint of low direct unit costs, but it could have higher absenteeism, and turnover, as well as a tendency to greater errors. In such a department this initial unit cost efficiency should be readjusted (and it would be downward) because of the additional costs incurred due to the turnover, absenteeism, and correction of errors. The costs of turnover and absenteeism usually remain hidden in the various burden, administration, indirect, and similar

accounts. These costs do exist, and some departments waste much more money per employee than do others. Often a department can reduce its total unit costs more through concerted effort on these indirect costs than it can through merely increased output. However, such costs are usually spread on a per employee or per dollar of direct payroll basis so that the more efficient departments from the standpoint of these indirect charges have to carry the load of the more poorly run departments. Since such factors as turnover and absenteeism relate to both costs and morale, it should pay top management to reward those in charge of the departments who are able to hold these factors to a practical minimum.

Since the correlations between the six factors and morale tend to be somewhat higher than the correlations between all of the various factors, it paid to compute a multiple correlation. The multiple correlation was .71 which is high enough to warrant the use of these six factors for the construction of an objective morale index. Table 3 lists the Beta weights for each factor.

Table 3

Factor	Beta Weight
1. Per Cent Productive Efficiency	.0630
2. Per Cent Error Efficiency NOT Affecting Customers	.1674
3. Per Cent Error Efficiency Affecting Customers	.1227
4. Per Cent Turnover	.4348
5. Per Cent Late	.0758
6. Per Cent Absent	.4894

$R_{9.123456} = .706$ $r_{t.uvwxyz} = .706$

In order to check our calculations we multiplied each factor for each department by the Beta weight given in Table 3 and then added the six products for each department to arrive at an objective morale index. We correlated this objective morale index with the morale score and obtained .706 which is shown after $r_{t.uvwxyz}$ in Table 3.

Recommendations

Since this study has determined that there is an important relationship between morale and the combined factors of production efficiency, error efficiency, labor turnover, tardiness and absence it is recommended that the index of these factors be used as a determinant of the relative levels of morale in various operating departments. The Objective Morale Index is also a measure of supervisory effectiveness and may be used to supple-

ment or replace supervisory merit rating. Costing of the factors will enable management to measure the value of morale improvement efforts in each local situation.

Those departments falling into the lowest morale classifications should be carefully scrutinized to determine whether or not the problem is department-wide or if it is limited to certain activities or working units. Examination of the index factors of each group will determine this. When low morale has been localized as much as possible a diagnostic investigation should be conducted to determine causes. Corrective action is then to be applied where it is needed.

Another questionnaire survey should be undertaken to secure data for a re-validation of the Objective Morale Index. Thereafter the questionnaire survey need be used only at longer intervals and not so much for morale measurement as for obtaining information for the diagnostic investigations.

Summary

The purpose of this study was to predict the morale of departments from objective data. A morale questionnaire was scored so that a quantitative score was available. The six objective factors of per cent productive efficiency, per cent error efficiency *NOT* affecting customers, per cent errors efficiency affecting customers, per cent turnover, per cent late, and per cent absent were intercorrelated and correlated with the departmental morale score. The multiple correlation between these 6 factors and morale was .71.

Since the six factors are objective and in most cases cost accountable their use for morale measurement is not only practical, but meaningful to operating executives in business and industry.

The relative weightings for the six factors for the prediction of morale are probably specific to this particular business. Therefore, a similar analysis must be made before this method of morale measurement can be applied to other businesses or industries.

Since the economic milieu may have some effect on the relative weights of the factors, a study is now in progress to revalidate the Objective Morale Index.

Implementing an Employee Opinion Survey

Major Fred E. Holdrege, Jr., USAF

Air Materiel Command, Dayton, Ohio

The problem of securing whole-hearted support from supervisors and employees is basic to the proper conduct and "follow-through" of an employee opinion survey. It is believed that many readers of the *Journal of Applied Psychology* will be interested in the information used to implement the Air Materiel Command's 1948 survey of the opinions of approximately 80,000 civil service workers employed at its eight major bases located in various parts of the country.

The survey itself was listed as a "Survey of Employment Conditions" in a covering letter which accompanied the opinion questionnaire itself. The letter is as follows:

Subject: Survey of Employment Conditions

To: Civilian Employees of the Air Materiel Command

1. To provide an opportunity for you to express your sentiments regarding personnel policies and practices which affect yourself and your employment with the Air Materiel Command, I have directed the Chief, Personnel and Administration Department, this Headquarters, to conduct a survey by means of the inclosed questionnaire.

2. It is my hope that this survey will reveal any existing difficulties and discrepancies which will be corrected for the benefit of all employees by the proper application of personnel management techniques.

3. You can help to make this survey a success by basing your reply to each question on an honest estimate of your views as to the manner in which the AMC is operated, considering, of course, the restrictions imposed upon us by Department of the Air Force, Civil Service directives, and other legal restrictions.

J. T. McNarney
General, USAF
Commanding

Instructions for filling out the questionnaire as printed on the opinion questionnaire were as follows:

How I Feel About My Job
(A Survey of Civilian Employee Attitudes)

You are asked to answer this questionnaire as a part of a survey of civilian employee opinions and attitudes which is being made by the Personnel Analysis Office, Headquarters, Air Materiel Command. The purpose of this survey is to find out how our employees feel about their work and the conditions under which they work, so that sound plans for improvement can be made.

This is not a "test." There are no right or wrong answers. Just answer the questions in the way you, yourself, feel about them. That is the only correct answer.

Instructions

Read each question or statement carefully to make sure you understand it before marking your answer. If you have any questions raise your hand.

Mark only one answer to each question. If you have more to say, write it on the last page of this questionnaire, but first mark one of the suggested answers which most clearly expresses your opinion.

Before turning in your questionnaire, check to make sure you have answered every question.

Be completely honest in your answers. That is the only way you can make this survey helpful to yourself and to the Command.

Do not write your name on the questionnaire.

Important: This questionnaire must be returned before you leave the room.

The questionnaire itself was 24 pages long containing a total of 146 items calling for multiple choice checking of answers and open-end answers as well as a page for making further remarks if desired. The contents may be classified as follows: Part 1, My Job, 25 questions; Part II, Personal History, 8 questions; Part III, Job Relations, 29 questions; Part IV, Supervision, 29 questions; Part V, Training, 10 questions; Part VI, Employee Relations, 18 questions; Part VII, Employee Welfare and Services, 13 questions; and Part VIII, Working Conditions, 14 questions.

The responses were punched into IBM cards and a detailed statistical report was prepared for each installation, copies being sent to the Commanding officers of each installation. Each report was accompanied by a Foreword designed to secure maximum attention and maximum action on the findings. The Foreword is as follows:

1. *Purpose:* This final report has been prepared to facilitate ready comparison of the findings at your installation with that of the Command average on the results of the employee opinion survey conducted at major AMC installations, March–April 1948.

2. *Value:*

a. In the final analysis, the value to be derived from this opinion survey depends upon whether management is psychologically prepared to translate the findings of the survey into action. Without action, the survey is merely a scrap of paper. With action, the survey becomes an effective means to an end . . . that of improved Command operational teamwork and economy.

b. It has been found that top management is too often inclined to take action only on incidental or secondary findings of such surveys . . . ignoring the more important revelations. In some instances, top management is merely content to learn that the morale of its organization is no worse than that of others.

c. Naturally, there is a defensive reaction on the part of operating officials to adverse disclosures brought about through morale or opinion surveys due to the interpretation of these disclosures as a reflection upon their ability. Hence, there is a tendency toward unwarranted discounting of survey data and a closing of the mind to self-analysis.

d. An objective viewpoint rather than a justification attitude must be taken if the results anticipated are to be gained.

3. *Employee Opinion:*

a. In reviewing the graphs and tabulations in this booklet, it is extremely important to bear in mind that these opinions of your employees are based on either (1) *factual conditions*, or (2) upon *erroneous impressions*. Whichever is the case, the only satisfactory solution lies in prompt remedial action.

b. It is also realized that adverse comment by employees with respect to a particular activity is not necessarily indicative of an activity poorly administered. It may be that from management's standpoint the activity is an efficient one. Whichever is the case, is of minor consequence; adverse employee opinion on any particular activity is a danger signal that management cannot afford to ignore. It is a sympton that calls for diagnosis and treatment.

c. If adverse opinions are based on factual conditions, then corrective action with respect to these conditions seems the only logical solution. If, on the other hand, unfavorable opinions expressed by employees are based upon erroneous impressions or information, then action should be taken to bring the true facts to the attention of employees so that they may be more thoroughly and accurately informed.

4. *Comparative Charts:* A Command comparison of the nature presented herein naturally places emphasis upon deviations from the average of the Command. These deviations are important, of course, but other aspects of the graphs must not be entirely disregarded. For example, matching or equaling the Command average is not necessarily indicative of a satisfactory condition for the reason that the Command average may reflect a condition unfavorable throughout the entire Command. Then too, it should be pointed out that the Command average is a composite picture of all installations and that a proportion of the installations must necessarily fall above the Command average.

5. *Chart Interpretation:* In order to assist you in interpreting the charts, recommendations have been made on the survey questions which show a significant variation from the Command average. While such recommendations naturally cannot be as accurate as on-the-spot observations, they are a composite of the thinking of various AMC Headquarters' organizations coupled with observations derived from the questionnaires themselves.

6. *Assistance:*

a. In order to assist those installations which fell below the average of the Command on various questions, names of the installations which presented the most favorable picture have been set forth in the final section of this report. Installations below the Command average are urged to contact these "high" installations for suggestions as to possible means of improvement. Liaison with these installations, leading the Command in methods and procedures used, is encouraged. An exchange of ideas is certainly one source of self improvement.

b. If assistance of the Hq Personnel Analysis Office is desired with respect to any morale or employee attitude problem, a brief statement of the problem should be forwarded to this Hq, Attention MCAA. If conditions warrant, an on-the-spot analysis will be made.

7. *Data:*

a. It should be remembered that the data secured in the conduct of this survey represent the first tangible information available to management as to what employees think of Command management and operations. Management can estimate, have a fairly good idea, or may be quite certain it knows pretty well how employees think . . . but up until the time of the survey this has been but a guess.

b. Management's opinion of Command operations, on the other hand, is fairly well-informed through the media of Inspector General reports, liaison between Command officials, Comptroller reports, etc. You might ask, "Just how important is it for management to have a thorough understanding of employee opinion?" Actually, having this one understanding is an absolute prerequisite to successful management. Management cannot afford to forget that "Management is the development of people and not the direction of things." Every policy that is written, every plan that is developed, every decision that is made, and every activity that is initiated must be considered in terms of the capacity of people to make them successful.

8. *The Employee:*

a. This seems an appropriate place to call attention to the one fellow who may be overlooked in follow-up of the questionnaire . . . and that is the employee who volunteered the information. He has a right to expect management to inform him of any action taken on the information he supplied. Correspondence reaching this Headquarters indicates that there is still much to be desired from an employee standpoint as to action being taken. It is realized that some adverse situations cannot be solved over night, nor even in months, yet unless the individual employee is acquainted with steps being taken or under consideration, he is likely to jump at the conclusion that his opinion is not even being considered. The failure of management to "keep faith" will nullify any further attempts to gain the cooperation of employees.

b. It cannot be over emphasized that employees must be informed of what is being done. They fulfilled their part of the contract by giving their honest opinion and suggestions for the improvement of working conditions. In General McNarney's letter to each employee, it was pointed out that adverse situations revealed by the survey would be corrected for the benefit of all employees. Thus, information on questionnaire follow-up action must be passed on to employees if the management-employee relationship is to remain unimpaired. Various media can be used to inform employees, such as notices, articles in the installation newspaper, supervisors' meetings, etc.

9. *Action Taken:* This Headquarters would also like to know of action taken on survey results. Communications should be addressed to the attention of the Personnel Analysis Office (MCAA). An analysis of follow-up action will be made and those steps which appear applicable to other field installations will be disseminated for the information and assistance of all.

10. *Results to be Expected:*[1] When results of the morale survey are properly accepted and acted upon, the following benefits may be expected:

a. The possibility of a permanently higher level of employee morale and productivity brought about by concrete change in conditions, practices, and policies.

b. The improved morale brought about by the very administration of the survey itself. Most surveys bring about immediate morale gains based upon the employees' discovery that management is really concerned about their feelings. When changes are initiated as a result of employees' opinion and it is made known to them that their feelings have caused the change, they will feel that they have some part in the determination of management policies. If expected changes are not forthcoming or long delayed, the temporary boost in morale is soon lost as a result of employee disillusionment and frustration.

c. The opportunity to increase the understanding of management employee

[1] *Making the Most of Morale Surveys* by F. F. Bradshaw and Herbert E. Krugman of Richardson, Bellows, Henry and Company, Inc.

problems by experimenting with new policies on a small scale or by intensively studying problems raised by the survey.

d. If action is taken as a result of an initial survey, employees will be more voluble in their replies to questions on which they are asked to comment in future surveys; hence, more effective results can be obtained from these questions on which employees are asked to comment.

e. An opportunity to commit the organization to a permanent goal of ever-rising morale levels which can be measured through the means of annual morale surveys.

In addition to the Foreword each report also contained two sample letters together with a suggestion as to how they might best be used. The suggestion for the preparation and use of the follow-up letters is as follows:

Sample Letters of Employee Opinion Follow-Up

As an illustration of one means of following-up on the opinion survey, the following pages contain two suggested survey action letters which may be easily adapted to the local situation by Commanding Generals of the various Air Materiel Areas.

One is a letter to all employees acquainting them with remedial action taken on adverse survey disclosures; the other a letter to all supervisors pointing out supervisory functions in need of improvement.

The first sample letter intended for employees is as follows:.

<div align="center">Headquarters
Air Materiel Area</div>

To all employees:

I would like very much to be able to talk with you individually to obtain your attitudes and opinions concerning the conditions under which you work each day. Since this is not possible, I welcomed the opportunity to have the AMC employee opinion survey conducted at this Headquarters. This appraisal of all phases of your work is of particular interest to me as an effective means for improving working conditions, morale, job satisfaction, and work simplification.

In view of the time involved in determining the practicability of making suggested changes and the planning and working out of necessary readjustments, many of the improvements which you have proposed cannot be put into effect immediately. However, the following action has already been taken as a result of your suggestions:

1. Exhaust fans have been installed in the aircraft repair shop.

2. Fluorescent bulbs have been provided for all light fixtures in the instrument repair shop.

3. Equipment and tools in the repair shops have been rearranged to eliminate blocking of exits.

The following projects are presently being given special consideration:

1. Survey of the procurement and routing of parts to expedite delivery.

2. Plan for improving parking facilities and improving traffic controls.

3. Survey of the wage scales in this locality with the intent of adjusting prevailing pay rates and publishing for employee information an approved wage schedule.

I desire to express my appreciation of the many opinions and comments submitted. It is my hope that any unfavorable situations will be corrected to the greatest extent

possible. Our combined efforts to improve working conditions and promote good relations will provide all of us with a still better place in which to work.

                                    Commanding General

The second sample letter intended for supervisors is as follows:

Headquarters
Air Materiel Area

To all supervisors:

An impartial analysis of the AMC employee opinion survey encompassing every phase of our work activity, has afforded me a wealth of information that I have accepted as an accurate diagnosis of our "state of health." We, as management, must assume the responsibility for the unfavorable conditions in our personnel relations as well as the favorable aspects presented by our people. However, *you* are management to the employee and the manner in which you execute the duties and responsibilities necessarily delegated to your position influences the employee's attitude toward his employment and the entire management program.

Supervisory training courses equip you with the fundamentals or principles of supervision but this knowledge alone is not enough to ensure effective practices. To attain skill in the art of supervising people, just as in playing golf or driving a car, requires diligent practice and broad experience in order to successfully cope with situations that do not fall into normal standard patterns. In addition to acquiring a knowledge of the basic principles of supervision and a degree of skill in supervising people, one must be able to identify and "think through" the elements of his job and then analyze his managerial practices objectively. As a result of such analysis you may decide that you are uncertain as to what your actual responsibilities are, or perhaps you do not have a thorough knowledge of pertinent rules and regulations or have unconsciously drifted into some poor habits of thinking and doing. The logical conclusion is the realization that you want to acquire more skill in supervising the people in your group. Having reached this conclusion, you are in the frame of mind to welcome ideas, information, and instruction that will help correct those phases of supervision in which you have rated yourself weak.

It is to this state of open-mindedness that I appeal in presenting to you now and in the future my philosophy of effective supervision and personnel relations practices. You are not expected to accomplish this analysis or program for improvement alone but through the combined efforts of our management team we can strengthen our weaknesses.

In order to give you a background for this program, I have selected for brief discussion a few supervisory responsibilities that have been neglected to varying degrees.

1. A significant number of employees receive direct orders from two or more people. This practice is not only in violation of sound management principles, but evidence of indifference toward our organizational structure. All organization charts clearly define lines of authority and control and have been planned to ensure both work and employee efficiency at even the lowest levels of operation. When we permit these lines to become crossed or tangled, we are short-circuiting our work efficiency and inviting personnel problems.

2. Giving an employee the credit he deserves for doing a good job or making a good suggestion requires very little time and effort but has a tremendous effect on groups attitude. You have nothing to lose in giving credit or recognition for it is merely

passing on to another that which rightfully belongs to him. The practice of giving credit where and when it is due is one of the most impressive "tools" of supervision for the craving to be appreciated and recognized as an individual of worth is a universal human trait.

3. We are by nature creatures of habit. It is a natural tendency to resist changes or to discredit the changes that another has proposed. You must reach beyond this narrow view and welcome all suggestions intended to improve or streamline operation. Give each suggestion serious consideration; if it is not advisable to make the change, explain your reasons (which must be sound) to the employee. If the suggestion has merit and can be incorporated in the operation, give the employee credit for proposing it. The attitude you assume in the initial discussion of the suggestion and the manner in which you receive ideas will either make employees feel free to bring other ideas to you or discourage them from ever approaching you again. We must encourage constructive thinking. It was evidenced in the Survey analysis that employees want to be happy in their work—it is our responsibility to provide the opportunity.

4. It is my contention that at least half of the dissatisfaction and complaints can be attributed to erroneous information or to the lack of dissemination of information to employees. A good portion of the blame for this falls on our shoulders for many of the criticisms received concerned items which a good supervisor could answer. As an example, a surprising number of employees do not have a clear concept of the difference in the work performed by the Civilian Classification Branch and the Civilian Utilization Branch. These are two distinct personnel services and every employee at some time or another is influenced by decisions made by these Branches. This type of information should be common knowledge among employees. A supervisor who shares information stimulates a good group spirit and gives the individual an assurance of some prestige in the group.

It is you, the key men and women on the management team, who have been entrusted with the grave responsibility of effectively applying the principles of supervision. Clear your minds and thoughts of prejudices or selfish ambitions; be eager to accept new ideas and sound procedures. The degree of your cooperation will determine the success of our personnel program.

....
....
Commanding General

Information obtained from the survey has proven its value even beyond that originally anticipated. As a result of the questionnaire, various phases of the over-all "working conditions" picture have been subjected to re-evaluation in light of employee opinion.

Many improvements have been made. Of course, there are a number of situations which will require time and continuous remedial effort before some degree of satisfaction can be assumed. On those problems extremely complex, such as the proper utilization of skills and abilities, progress can be achieved only by constant surveillance and study. Then too, we must be extremely careful to insure the maintenance of standards on those practices most favorably rated by employees.

Plans are now going ahead for the conduct of the second command-wide survey. With the results of the first survey to serve as a basis for

comparison, it will be possible for the first time to measure improvement of existing policies and practices from an employee standpoint.

An attempt at a "one-shot" approach to achieve a "once-and-for-all" solution is being carefully avoided.   Instead, the picture of the Command's operations as presented by employees will serve as a guide and a basis for the checking and development of a personnel program designed to achieve maximum accomplishment through our human resources.

A Trade Test for Power Sewing Machine Operators

Edward Glanz

Teachers College, Columbia University

The David Clark Company Inc., like many other garment factories, hires many power sewing machine operators throughout the year. The problem has been to reduce the turnover resulting from having to release unsatisfactory stitchers, and also to free foreladies from having to teach new employees how to operate a power sewing machine. In the past reliance was put on the statements of the applicants as to their skill and experience. This process frequently yielded poor workers and was expensive in turnover and foreladies' time.

The problem at David Clark Company Inc., a subsidiary of Munsingwear Inc., was to find out which of the numerous applicants claiming skill and experience could actually operate a power sewing machine on the lightweight cotton used in making brassieres and girdles.

Design and Sampling

In attempting to devise a trade test for stitchers pertinent previous work was carefully studied [1] and production line work was observed. The test that is now to be described is based on an hypothesis resulting from a combination of these two sources and validated on present employees. The four sections of the trade test as well as the trial sample were given to all of the power sewing machine operators in the plant. These operators were unselected as to level of skill or productivity except as experience itself is a selector. There were forty-nine operators, all female. Their experience ranged from two months to ten years and their productivity ranged from a barely acceptable amount of work to twice the acceptable amount of work.

The plan was to correlate the results of the trade test with the supervisors' ratings and production records. If validated, the trade test could then be used to classify applicants on a proficiency scale.

Criteria Selected

Supervisors' ratings were obtained from the two general supervisors of all of the operators. Both of these supervisors were qualified rate setters

---

[1] See especially: Otis, J. L., The prediction of success in power sewing machine operating. *J. appl. Psychol.*, 1938, 22, 350–366. Blum, M. L., Selection of sewing machine operators. *J. appl. Psychol.*, 1943, 27, 35–40.

and were experienced in evaluating the speed and quality of the operators' work.   The ratings were obtained independently and before any test results were available.   The speed and quality ratings were obtained for each worker and combined into an overall rating.   The product moment correlation of the two supervisors' overall rating agreement was + .87 ±.035.   Each operator was rated on the following five point scales for speed and quality:

Speed Rating Scale	*Quality Rating Scale*
Very Fast Worker	Highest Quality Work
Faster Than Most	Quality Better Than Most
Average Speed Worker	Average Quality Work
Slower Than Most	Quality Poorer Than Most
Very Slow Worker	Very Poor Quality Work

Production records were obtained for a one year period in order to provide a further check on the trade test.   The product moment correlation between these criteria was + .83±.043.

The Trade Test

Trial Sample.   The trial sample consisted of a single piece of lightweight nude cotton 5¾″ by 7¾″.   The subject is instructed to stitch around the cloth approximately a quarter inch from the edge.
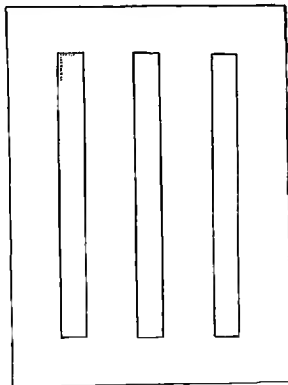
Fig. 1. Worksample 1.

1. *Bias Tape on Cotton.* The first worksample consisted of a piece of lightweight nude cotton 5½″ by 8½″ onto which three pieces of white cotton bias tape 6½″ by ½″ are to be sewed. The subject is instructed to sew these strips onto the cotton cloth as straight as possible and to space them as well as possible without a ruler.

2. *Hemming Material.* The second worksample consisted of a piece of lightweight nude cotton 5¾″ by 7¾″. The subject is instructed to sew a double lap hem into the material, turning a quarter inch under each time.
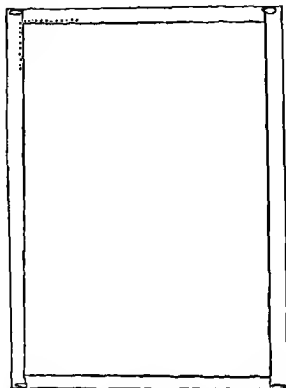
Fig. 2. Worksample 2.

3. *Stitching Between the Lines.* The third worksample consisted of a piece of medium thickness cardboard 5½″ by 8½″ with a double lined pattern to follow. The subject is instructed to follow the pattern drawn on the cardboard stitching without thread between the lines without going outside the lines.

4. *Stitching on the Line.*[2] The fourth worksample consisted of a piece of medium thickness cardboard (same as in worksample number three) 5½″ by 8½″ with a single lined pattern to follow. The subject is instructed to follow the pattern drawn on the cardboard stitching without thread and without going off the line.

[2] The seeming similarity of Worksample numbers three and four is only superficial for number four is actually much more difficult.
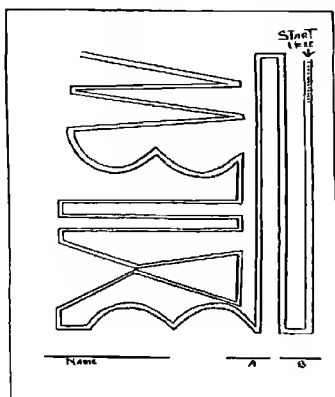
Fig. 3. Worksample 3.
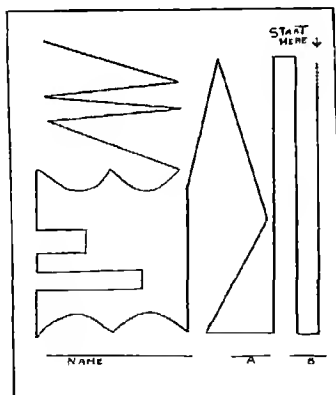
Fig. 4. Worksample 4.

Administration of the Test

The administration of the trade test is carried out by the use of shallow boxes lettered SAMPLE 1, 2, 3, 4. The material the subject is to sew with is placed inside the box along with an example done accurately and carefully so that the subject may see exactly what is to be done. Typewritten instructions on the inside bottom of the box tell the subject what to do. This method of utilizing both written and visual instruction insures that persons of all types may be tested accurately and also provides a uniform administration procedure.

The boxes are given to the subject one at a time and the administrator instructs the subject to do the task neatly and accurately, but also quickly. The same instructions are given to operators on the production line. The time of each operation is taken from the moment the box is placed on the sewing machine until the operator has trimmed the threads or until the machine is stopped on the cardboard.

Scoring the Test

The trade test is scored for both speed and quality and also for a combination of these which gives an overall (summed) stitching score.[3] Since each test is timed, the total time of each is added thus giving a speed score for the battery. The quality scoring is more complicated for many things must be taken into consideration. Worksample 1 and worksample 2 were scored by setting distance limits for each operation and by counting incorrect stitches. In number 1 placement, parrallelism, and evenness were measured objectively with a ruler and the incorrect stitches counted. Handling, folding, and stitching were scored in the same way in number 2. The time needed to complete number 2 was also figured in the quality score by adding to the score for quickness and lowering the score for slowness, for the better stitchers can fold and handle the material more skillfully than the poorer operators. The quality scoring of worksamples 3 and 4 was much simpler: the number of holes that were outside the lines in number 3 and the number of holes that did not touch in number 4 gave the quality score.

This scoring procedure is based on objective measurements and counting of stitches and thus can be carried on by others very easily.[4]

[3] The advisability of combining these separate and distinct ratings into an overall stitching score may be questioned since the two scores are more or less opposed to each other: that is, if the time or speed rating is high the quality rating is apt to be low and vice-versa. However, an accurate picture of the stitcher's ability cannot be obtained many times without combining these two scores because of the above fact. Of course, it must always be added that the overall score is made up of the quality and speed rating.

[4] The norms developed for this trade test, as well as the scoring method, are available for use, but local norms should be developed by industrial users.

Results

These results are clearly significant. The supervisors' ratings yielded higher correlations probably because the supervisors knew the operators' work and could correct for the distortion that appeared in the production records as explained in the footnote to Table I. The two sets of correla-

Table 1
Trade Test and Criteria Correlations

	Correlations	Standard Error
1. Trade Test and Supervisors' Ratings:		
Speed Scores	.58	±.095
Quality Scores	.55	±.090
Overall Scores (Summed)	.64	±.084
2. Trade Test and Production Records:		
Speed Scores	.56	±.097
Quality Scores	.31	±.120
Overall Scores (Summed)	.53	±.103
3. Trade Test and Combined Criteria:*		
Overall Scores and Combined Criteria (Summed)	.67	±.078

* Combined by means of Z scores and also weighted in a 2-1 ratio, supervisors' ratings to production records. This is because many operators of only mediocre skill become accustomed to a single task and accumulate rather high earnings over a period of time. This may also explain the somewhat lower correlations obtained with production records.

tions and the combined criteria correlation do establish the trade test as an instrument that will differentiate between the more highly skilled and productive operators and the poorer and less productive operators.

Summary

The trade test for power sewing machine operators provides a method by which the more highly skilled and productive operators may be differentiated from the less skilled and less productive operators.

Such a test should be useful in selecting applicants who have the skills needed in production and who, consequently will require a minimum of on-the-job training.

Prediction of Job Success from the Application Blank

Willard A. Kerr

Illinois Institute of Technology

and

H. L. Martin

Radio Corporation of America

While considerable factual information is contained on the typical industrial personnel application blank, little information is now available to indicate the actual value of this information for predicting the probable job success of the applicant. This study attempts to make a small contribution to existing knowledge on this topic by obtaining correlations between success on the job and such information items as sex, marital status, possession of telephone, street address (i.e. part of city), age, birthplace (in or out of state), children, dependents, height, weight, previous employment with company, insurance, recent illness or operations, number of personal references listed, organizations, hobbies, company acquaintances, education, and previous positions for 244 employees in the personnel, engineering, purchasing, production control, phonograph record manufacturing, electronic tube manufacturing, and warehouse departments of the Indianapolis plant of the RCA Victor Division of Radio Corporation of America.

Success on the job measures for these 244 employees were obtained from supervisors and the raw merit ratings (split half reliability of the merit rating form was found to be above .75) from each supervisor were transformed into standard dichotomous scores which were then plotted with the information items to obtain tetrachoric coefficients of correlation between job success and these variables. These are presented in Table 1. Correlations significant at the five percent level or better are set in boldface type.

On the basis of the highest correlations, eleven items were scored, check-list fashion, and the total scores were correlated with job success to obtain a coefficient of $.35 \pm .04$.

Although all these findings should be accepted as tentative, it is possible that some of the findings may be found to apply to most departments of work in general industry. Analysis of item predictive value for various types of work was not attempted here because of the limited number of cases.

Marital status has the greatest relationship with the criterion for these cases. Area B street address correlates positively with job success while Area A address correlates negatively; this is regarded as surprising since the large area of above-average socioeconomic status in the city is in Area A. Number of children and number of dependents, while regarded as important by most personnel workers for obvious social reasons, do not correlate significantly with job success. Height and weight seem to be of little general predictive value, although obesity tends to be a definite

Table 1

Correlations between Job Success Ratings and Personal History Items

Female sex	−.16
Marital status: single	−.18
married	.30
divorced	−.05
Telephone number (possession of)	.07
Street address: Area A	−.22
Area B	.23
Area C	.15
Area D	−.11
Age	.08
Birthplace (in same state as plant)	.15
Number of children	.00
Number of dependents	.00
Height of males	−.12
Height of females	.05
Weight of males	−.27
Former employee of same company	.22
Holds insurance policy	.06
Recent illness or operation	.00
Number of personal references listed	−.17
Number of organizations in which membership is held	.23
Number of hobbies	−.18
Number of company acquaintances	−.09
Education: special training	.15
college	−.01
Number of previous positions	−.22

handicap for men. Former employment in the company seems to be an asset, but possession of insurance and recent illness or operation appear relatively unrelated. Listing of an excessive number of personal references, hobbies, or previous positions is negatively related with the criterion, but membership in organizations and special education are positively related.

It should be emphasized that these correlations are low and the findings possibly may apply only to the workers measured. Nevertheless, it is interesting to note that approximately ten per cent of the variance

in job success of these 244 employees is accounted for by "autobiographical" factors reported in the original applications for employment. Such check-list autobiographical scores may make a highly useful addition to the total predictive test battery. Naturally they should not be weighted more heavily in determining selection than their relative contribution to determination of job success variance indicates. In order to maintain the validity of the autobiographical scoring key, it should be revised periodically according to results of routine revalidations.

Better results may be obtained in selecting for a specific job with this device than when using it to hire for the entire plant. Manson (1), for example, found a coefficient of correlation of .40 between the weighted scores on an application blank and the production records of life insurance salesmen, and Ohmann (2) obtained a correlation of .67 between his blank and the earnings of paint salesmen.

Summary

1. Most of the items on a typical industrial personnel application blank are easily quantified in check-list fashion on the basis of a previous item validation study against a job success criterion.

2. In this study, when the original applications of 244 employees were scored check-list (unweighted) fashion with a validated key, the check-list raw scores were found to correlate .35 with the supervisory merit ratings of job success.

3. Since in this study the application blank accounts for approximately ten per cent of the variance in job success of an extremely heterogeneous (almost "run of the employment office") group of employees, it seems reasonable that the application blank or a systematic autobiographical inventory should become a standard part of the psychometric battery in industry.

4. In view of the facts that background factors change in predictive significance with time and their significance also is altered by changes in the business cycle, the industrial psychologist should revalidate such an instrument periodically.

5. When validation keys are developed for specific kinds of employees or job families, more substantial correlations are likely to be obtained both with the job success criterion and the tenure criterion.

References

1. Manson, G. E. What the application blank can tell. *J. Personnel Res.*, 1925, **4**, 1–28.
2. Ohmann, O. A. A report on selection of salesmen at the Tremco Manufacturing Company. *J. appl. Psychol.*, 1941, **25**, 18–29.

Tests Used by United States Air Carriers

Nicholas C. Feronte

Marquette University

The use of tests in the selection and promotion of employees is coming into increasing favor with United States industry. The first marked interest in such tests resulted from their use by the United States Army in World War I. The extensive and successful use of tests by our armed forces in World War II has given added impetus to the use of tests in industry. In some areas personnel testing has aroused an almost phenomenal interest among business executives.

A survey of the various psychological tests used by United States Air Carriers (Scheduled and Certificated) is provided by the results of a questionnaire received from 24 companies. Cooperation in replying to the questionnaire was very gratifying and personnel directors of many of the companies indicated they would like to receive a copy of the results: A questionnaire was sent on July 14, 1948, to all companies listed [1] asking the personnel director of each company to indicate which tests were most useful in selecting and promoting personnel within his company.

Of the 24 companies that returned questionnaires 13 indicated extensive use of tests, 2 use tests only to select pilots, 2 stated they had not begun operation, 1 disclosed tests were not used and would not be used until there are at least ten applicants per job, and 6 indicated they were small organizations, hence depended upon the ability of management to determine the qualifications of those seeking employment. However, they have felt the need of using more scientific methods.

The questionnaire listed 50 tests selected from the article published by Louttit and Browne.[2] The tests were chosen on the basis of published reports pertaining to the validity and reliability of each. Further evidence of suitability based on the author's use of the tests and the comments found in Buros' book.[3]

[1] American Aviation Directory, Spring–Summer 1947, American Aviation Publication, Washington, D. C. It lists thirty-six companies; however, five are operated by companies included in the mentioned list.

[2] Louttit, C. M., and Browne, C. G. The use of psychometric instruments in psychological clinics. *J. consult. Psychol.*, 1947, 11, 49–54.

[3] Buros, O. K. *The Nineteen Forty mental measurements yearbook.* Highland Park, New Jersey, The Mental Measurements Yearbook, 1941.

Table 1

Tests Reported as Being Used Most Often

Type and Name of Test	Number of Times Listed
General Ability	
American Council on Education	1
California Job Test	1
x General Aptitude	1
Ohio State University Psychological Test	1
Otis, Self-Administering	10
Personnel (Wonderlic)	6
Wechsler-Bellevue	1
Achievement	
x Arithmetic	1
Nelson Denny Reading	1
x Special Knowledge	1
Mechanical Ability	
Bennett, Mechanical Comprehension	5
MacQuarrie, Mechanical Ability	1
x Manual Dexterity	1
x Master Mechanic	1
x Mathematics-Physic	1
x Mechanical Comprehensive	1
Minnesota, Mechanical Assembly	1
Minnesota, Paper Form Board	3
O'Connor, Wiggly Block	1
x Second Officer Selection	1
Stenquist, Mechanical Aptitude	3
x Technical Information	1
x Trade Test	1
Clerical	
x Filing	1
Gregg Stenography	2
Gregg Typing	2
Minnesota, Vocational Test for Clerical Workers	3
Thurstone, Examination in Clerical Work	2
x Stenography	4
x Typing	4
Special	
Vision Test Administered by a Physician	1
Interest	
Kuder, Preference Record	3
x Mechanical	1
Strong, Vocational	1
Personality	
Humm-Wadsworth Temperament	3
x Memory Test	4
Minnesota Multiphasic Personality Inventory	6

Code: x = Self Developed Test.

Only those tests reported as being used are listed. See Table 1. Blank spaces were provided on the questionnaire and the personnel director was asked to list such tests in use by his company which did not appear on my questionnaire. This instruction produced a rather lengthy list of additional tests. Other instructions on the questionnaire were as follows:

How long have you used tests?

If you have used any tests and have abandoned the practice, will you please name the test.

For the selection or promotion of what types of employees have you used tests?

Who administers your test selection and administration program?

Table 2

Type of Employees Selected or Promoted by Tests

Type of Employees	Number of Times Listed
Apprentice	8
Semi-skilled workers	7
Salesmen	6
Clerical employees	10
Unskilled workers	5
Navigators	1
Skilled workers	8
Pilots	8
Foremen and/or supervisors	6
Executives	2
Sales agents	1
Flight engineers	1
Stewardess	2

An examination of the results in Table 1 indicates that the standardized tests used most frequently are the Otis, Self Administering, Personnel (Wonderlic), Minnesota Multiphasic Personality Inventory, Humm-Wadsworth Temperament, Bennett Mechanical Comprehension, Minnesota Clerical Aptitude Test, Stenquist Mechanical Aptitude, and Minnesota Paper Form Board. The self developed tests listed most often are typing, stenography, and memory test. Practically all companies reporting administer an intelligence test and also a clerical test.

Table 2 lists the type of employees selected or promoted by use of tests. To date, little use is made of tests to select or promote sales agents. In general tests are used mostly to select office clerks, pilots, apprentices, skilled workers, and semi-skilled workers.

The survey discloses that a few companies inaugurated testing only two years ago, while others reported having used tests to select and promote personnel for the last ten years. It is interesting to learn from the questionnaire that no company discontinued administering tests permanently once it began to use them.

In replying to the question who administers your test selection and administration program? Six listed the employment manager, four delegated the assignment to the personnel clerk, three indicated the responsibility was assumed by the personnel manager, and one stated a clinical psychologist is employed full time to administer tests.

Perhaps the most interesting implication of this survey is that in general all air carrier companies are either using tests or have felt the need of using scientific methods for selecting and promoting personnel. Of equal significance is the fact that once a company inaugurated a testing program it was never halted permanently by management. The various company's interest in tests is evidenced by the use of a diversity and variety of not only well known standardized psychological tests but also company developed tests.

In general, companies reported using tests that have been found to be valid and reliable.

It would be gratifying to flight passengers to learn that the survey disclosed psychological instruments were used most often, in addition to selecting clerical employees, to select personnel definitely responsible for the maintenance and the flight operation of the airplane.

A Factor Study of Worker Characteristics *

Nathan Jaspen

Pennsylvania State College

In order to make the relationships between occupations more understandable, the Occupational Analysis Division of the United States Employment Service has selected the most significant job characteristics, and assembled them into a rating form adapted after Viteles' Job Psychograph (15, 16). This rating form has been applied to several thousand occupations. Estimates of the most significant worker characteristics required for each occupation are made independently by several trained analysts. If, for example, an assembly job demands an unusual amount of finger dexterity, the analyst indicates that dexterity of fingers is an important worker characteristic for this occupation. This information is punched on Speed Sort cards, which can then be sorted so that the occupations which have various characteristics in common can be studied and the relationship between them noted. The traits included in the Worker Characteristics Form are listed in Table 1.

The Worker Characteristics Form includes 45 traits or abilities which may be needed by the worker to do the job. A large number of important "Job Families," containing lists of occupations related to a single occupation or to a limited number of selected occupations, have been established on the basis of worker characteristics. These have had various uses: to select workers for critical occupations from related occupations; to transfer workers from occupations in which there were labor surpluses; to upgrade employees on the job; and, to show the civilian occupations related to military occupations (8, p. 703). Nevertheless, it is obvious that the usefulness of the Form for some purposes would be increased if it contained a smaller number of independent traits. This pilot study was undertaken to determine what basic factors were being measured

Table 1

Proportional Frequency with Which Traits in the Worker Characteristics Form Are Rated as Required in Significant Degree in 275 Selected Occupations in the Skilled, Semiskilled, and Unskilled Categories of Occupations

Characteristic Required of Worker	Per Cent
1.* Work rapidly for long periods	17
2.* Strength of hands	37
3.* Strength of arms	47
4.* Strength of back	27
5.* Strength of legs	11
6.* Dexterity of fingers	24
7.* Dexterity of hands and arms	48
8. Dexterity of foot and leg	05
9.* Eye-hand coordination	52
10. Foot-hand-eye coordination	05
11.* Coordination of independent movements of both hands	11
12.* Estimate size of objects	11
13. Estimate quantity of objects	06
14.* Perceive form of objects	21
15. Estimate speed of moving objects	04
16.* Keenness of vision	25
17. Keenness of hearing	01
18. Sense of smell	01
19. Sense of taste	**
20. Touch discrimination	07
21.* "Muscular" discrimination	15
22.* Memory for details (things)	16
23. Memory for ideas (abstract)	04
24. Memory for oral directions	04
25. Memory for written directions	02
26. Arithmetic computation	04
27.* Intelligence	09
28. Adaptability	04
29. Ability to make decisions	08
30. Ability to plan	08
31. Initiative	05
32.* Understanding of mechanical devices	15
33.* Attention to many items	16
34. Oral expression	01
35. Skill in written expression	**
36. Tact in dealing with people	04
37. Memory of names and persons	**
38. Personal appearance	01
39. Concentration amidst distractions	03
40. Emotional stability	05

Table 1 *(Continued)*

Characteristic Required of Worker	Per Cent
41.* Work under hazardous conditions	24
42.* Estimate quality of objects	09
43.* Work under unpleasant physical conditions	29
44. Color discrimination	07
45. Ability to meet and deal with public	02
Added Characteristics	
46.* Tools used	68
47.* Knowledge of graphic instructions required	17

		Per Cent
* Skill Level:	Skilled	38
	Semiskilled	41
	Unskilled	21

* Characteristics which are included in this study.

** Less than one per cent.

by the Worker Characteristics Form. As it developed, only 20 of the 45 traits included in the Form were included in this study, so the factors discovered have reference only to these 20 traits and not to the Form as a whole. However, this is not a serious restriction, as none of the remaining 25 traits was present in significant amount in as many as 10% of a sample of the occupations so far studied.

Description of the Data

The Worker Characteristics Form provides for estimates of 45 traits, in an A, B, C, or 0 amount. The amounts designated by these letters are (10, p. 176–178): A. A very great amount of the trait, such as would be possessed by not more than 2 per cent of the general population; B. A distinctly above-average amount of the trait, less than that designated by A but more than that designated by C; C. An amount of the trait less than that possessed by the highest 30 per cent of the general population; and 0. The trait is not required for the job.

The analyst is instructed to compare the worker on the job with people in general, outside the plant and even outside the industry. The estimates are based on the abilities demanded by the job, not the abilities which the worker under observation happens to have. If the analyst is in doubt between an A or a B amount, he is instructed to rate the characteristic as B; if he is in doubt between a B or a C amount, he is supposed to rate the characteristic as B; and if he is in doubt between a C or an 0 amount, his instructions are to rate the characteristic as C (3, p. 2). In this way, the significant characteristics of the job are not submerged in the C column but are focused more clearly in the B column; A amounts of a trait are very definitely indicated; and traits which are necessary though in a small degree are not lost in the 0 column.

The analyst is guided in his understanding of the meaning of the traits by a manual embodying comprehensive definitions of the traits and examples of the different quantities of each trait (3). Several, usually ten, separate estimates are submitted by as many trained analysts who make their observations and estimates in different plants and states. The reliability of these estimates is not known, at least by the present writer. The information is collated by an analyst at headquarters, and reviewed by an analyst of higher grade. The final ratings are punched on Speed Sort cards, one for each occupation, in two amounts: A or B on the one hand, for traits which are present in significant amounts, and C or O, on the other, for traits which are not present in significant amounts. The frequency with which these traits were rated as significant in the study sample is shown (expressed in percentages) in Table 1.

About 9000 Speed Sort cards have been punched up to 1943. All kinds of jobs are included: professional, managerial, technical, service, sales, clerical, agricultural, skilled, semiskilled, and unskilled occupations. The present study has been restricted to the categories of skilled, semiskilled, and unskilled occupations.

The Speed Sort cards are arranged in sequence by occupational code. The skilled occupations are the 4-00 to 5-99 series; the semiskilled are the 6-00 to 7-99 series; and the unskilled are the 8-00 to 9-99 series. Two cards were selected at random from each centile group from 4-00 to 9-99. This would have yielded 1200 cards if the file had had sufficient cards in each centile code; but in many cases a centile code was open, or no studies had been conducted of occupations within the centile code, or only one study had been conducted, in which case only one card was selected. Two hundred and seventy-five Speed Sort cards were selected in this fashion from about 7500.

The occupational codes represent different occupations, and no account is taken of the relative number of workers in each occupation. Skilled jobs are more finely differentiated than unskilled jobs. There is a certain bias in the occupations selected by the United States Employment Service for study; the cooperativeness of the different industries, and the matter of geography, are only a few of the variable factors. No one knows how representative the 7500 Speed Sort cards are of skilled, semiskilled, and unskilled occupations in the United States. The 275 study occupations may be charitably regarded as a sample of the classification structure of a major part of the Dictionary of Occupational Titles (14).

Procedure

Obtaining the Correlations. Since each characteristic, as punched on the Speed Sort cards, was either present or not present, it was feasible to compute only tetrachoric correlation coefficients between pairs of characteristics. Thurstone recommends that the coefficient not be computed at all if one of the tail areas is less than 10 per cent of the total population (1). Consequently, all characteristics which were significantly present in less than 10 per cent, approximately, of the sample of jobs in the present study were eliminated, since valid correlations for them could not be computed. Only 20 of the worker characteristics survived this step. In addition, the variables (46) "Tools used," and (47) "Knowledge of graphic instructions required," remained. The twenty-third variable of the study was Skill.

Skill was expressed in three categories—skilled, semiskilled, and un-

Table 2

Matrix R_0. Correlation of 23 Characteristics for 275 Occupations *

Skill	1	2	3	4	5	6	7	9	11	12	14	16	21	22	27	32	33	41	42	43	46	47
Skill	-47	-02	-06	-18	03	28	23	24	25	42	52	48	28	51	44	37	46	05	10	-03	18	42
1		22	-05	00	00	20	00	18	18	-10	-21	03	-20	-42	-14	-02	08	-22	07	-14	-22	-26
2			82	50	62	-07	44	24	25	30	34	-06	22	17	-04	13	02	27	06	15	34	22
3				88	72	-40	23	02	12	-02	11	-03	11	-01	-16	10	06	34	-13	27	33	06
4					91	-29	-15	-14	-15	-08	-02	-21	-10	-19	-06	03	04	24	-18	24	21	-03
5						-21	03	-02	-21	08	-21	00	-01	-14	-16	-10	25	-46	-05	08	13	01
6							-08	35	62	44	55	67	66	33	14	-05	-07	09	-14	-42	23	34
7								81	62	00	45	53	20	16	-02	21	03	-02	-15	01	34	36
9									44	10	37	26	48	08	-13	07	28	-17	-26	-19	37	34
11										55	67	48	39	00	-14	11	08	10	22	10	10	26
12											66	24	18	21	20	13	04	-10	28	-12	29	19
14												33	36	14	00	21	26	-20	43	-22	48	41
16													43	14	05	21	-10	-21	08	-37	09	02
21														08	03	11	70	11	32	08	34	15
22															25	23	28	12	09	00	00	42
27																54	46	55	11	08	08	20
32																	55	08	10	-15	30	52
33																		55	-06	04	-12	25
41																			-06	64	12	09
42																				-14	05	08
43																					11	-02
46																						39

* All entries have been multiplied by 100 in order to eliminate the decimal point.

skilled—and was determined from the occupational code of each of the 275 occupations in the study sample. The intercorrelations between skill and the other 22 variables were found from 2X3-fold tables with unequally spaced intervals. Normal distributions and rectangular regression were assumed, each interval was assigned the value of its mean expressed as a deviate on a unit normal curve, and Pearson product-moment r's were computed. A correction for "broad categories" was then applied (4, pp. 167–171; 7, pp. 399–402).

The matrix of intercorrelations is shown in Table 2.

The Factor Analysis Procedure. Eight centroid factors were extracted from the matrix of intercorrelations, by Thurstone's Centroid method (13). The centroid matrix is shown in Table 3. Estimates of the communality were used in the diagonal cells of the correlation and the residual matrices (13, p. 89), the estimates being the highest coefficients in each column. In the successive extractions the characteristics were

Table 3

The Centroid Matrix F_o. Projections of Worker Characteristics Vectors on 8 Arbitrary Orthogonal Axes Determined by the Centroid Method *

	I	II	III	IV	V	VI	VII	VIII	h²
Skill	63	33	31	−27	12	23	−04	−18	78
1	−21	07	−36	33	−37	−19	−11	−13	49
2	58	−45	−34	18	−22	−19	−19	26	88
3	45	−78	−23	08	−25	−13	−11	07	97
4	23	−81	−16	−09	−39	14	16	−12	95
5	29	−71	−17	−10	−37	38	09	−12	93
6	28	63	−37	12	−06	27	18	−08	74
7	53	13	−41	39	38	−16	07	−03	79
9	53	33	−59	32	30	07	07	−20	98
.11	46	32	−26	51	08	08	−12	−15	69
12	46	11	−16	−42	18	15	−40	09	65
14	66	25	−29	−32	19	16	−12	26	83
16	38	51	−21	−23	−16	−14	−16	−24	63
21	47	18	−22	−30	20	−39	18	−10	63
22	49	21	60	14	−18	22	−15	22	82
27	33	17	50	−06	−13	−10	17	−21	49
32	45	23	28	05	−30	−13	36	11	58
33	44	24	45	27	−50	17	−16	−23	88
41	21	−54	36	15	30	−10	−24	−13	66
42	11	27	13	−33	−18	−39	−32	28	58
43	10	−50	31	20	37	−18	−17	−29	68
46	54	−16	−16	−11	20	−11	30	24	55
47	58	14	19	17	11	13	31	36	68

* All entries have been multiplied by 100 in order to eliminate the decimal point.

Table 4

The Transformation Matrix A *

	A	B	C	D	E	F	G	H
I	28	39	40	24	37	29	34	01
II	−69	32	27	−30	33	−05	−07	19
III	−34	64	−14	40	−53	−04	15	−07
IV	−16	22	−70	08	65	03	−25	−20
V	−50	−48	09	59	18	−33	23	−24
VI	19	17	−06	−17	02	−86	−08	−24
VII	−06	04	−49	−33	04	−03	83	29
VIII	−09	−11	−08	−45	−15	25	21	−85

* All entries have been multiplied by 100 in order to eliminate the decimal point.

Table 5

Rotated Factorial Matrix $V = F_rA$ *

	A	B	C	D	E	F	G	H
Skill	−11	49	52	28	05	−14	24	16
1	13	−06	−20	−28	30	22	−40	18
2	62	−06	10	00	27	52	−01	−27
3	83	−12	−02	14	40	03	−14	
4	91	−06	−15	−03	−20	08	14	00
5	91	01	−07	−05	−14	−11	08	03
6	−17	19	15	−36	60	−17	06	19
7	−09	−13	04	20	78	18	16	−03
9	−05	−10	16	05	91	−04	07	15
11	−09	21	02	09	78	01	−15	04
12	13	−07	73	15	02	−03	−05	−15
14	08	−02	64	−05	27	05	23	−20
16	−05	20	57	−07	27	20	−08	39
21	−05	−11	44	12	21	37	39	30
22	−11	77	10	11	−04	04	07	−27
27	−13	57	08	20	−11	14	27	30
32	−04	56	−04	−17	07	34	43	13
33	07	88	03	07	12	07	−18	19
41	14	02	−06	70	−13	05	−02	−16
42	−14	12	47	−09	−25	48	−10	−08
43	05	−06	−13	75	−06	01	−03	00
46	17	−10	13	03	17	25	55	−13
47	−11	37	−05	−03	25	09	52	−29

* All entries have been multiplied by 100 in order to eliminate the decimal point.

reflected by a method which maximized the amount of the total variance which was accounted for by each new factor (13, pp. 99–100).

After eight extractions, the extraction process was discontinued because some communalities appeared to be spuriously great, and in fact

the ninth centroid factor would have increased two of the communalities to more than unity. Neither Coomb's criterion (2) nor McNemar's criterion (6) for the number of factors appeared applicable. This may have been because the scores in this study were based on judgment rather than tests.

The arbitrary axes of the centroid matrix were then rotated by Thurstone's method of extended vectors (11) until a solution was obtained which for the most part satisfies the requirements of simple structure. The transformation matrix A is shown in Table 4, and the rotated factorial matrix is shown in Table 5. An effort was also made to keep the correlations between the primary factors as close to zero as possible. The range of these correlations is from .16 to −.14

Interpretation of Factors

The significant factor loadings are here considered to be those of .40 and above. Thurstone notes that "the naming of a factor cannot be made with confidence unless the projections are as large as .50 or .60 so that the factor accounts for a fourth or a third of the variance of a test" (12, p. 79) or measure. Loadings below .20 are of no significance. Loadings between .30 and .40 may be important.

Factor A. The characteristics which enter significantly into Factor A, and their loadings, are: 4. Strength of back, .91; 5. Strength of legs, .91; 3. Strength of arms, .83; and 2. Strength of hands, .62.

Factor A has been designated as Strength. All of the loadings are phenomenally high. The intercorrelations also were very high, indicating the correspondence between the traits.

Factor B. The following characteristics have significant loadings on B: 33. Attention to many items, .88; 22. Memory for details (things), .77; 27. Intelligence, .57; 32. Understanding of mechanical devices, .56; Skill level, .49; and 47. Graphic instructions, .37.

The underlying factor in B appears to be Intelligence. The higher loadings of characteristics 33 and 22 may be due to a tendency on the part of the analysts to underestimate the amount of intelligence required for industrial jobs of a semi-clerical character.

Factor C. Factor C has the following loadings: 12. Estimate size of objects, .73; 14. Perceive form of objects, .64; 16. Keenness of vision, .57; Skill level, .52; .42. Estimate quality of objects, .47; and 21. "Muscular" discrimination, .44.

Factor C appears to be an Inspection factor, perhaps predominantly visual inspection. Characteristic 21 refers to kinesthetic sensitivity, and it and perhaps Skill level are the only characteristics here which are not visual:

Factor D. Factor D has significant loadings as follows: 43. Work under unpleasant physical conditions, .75; 41. Work under hazardous conditions, .70; and 6. Dexterity of fingers, −.36.

This factor, which may be nothing more than a doublet, has been designated Physically Unpleasant Working Conditions. The relatively large negative loading of Characteristic 6 may be a sampling error, or it may be attributed to the gross movements involved in work which is outstandingly physical.

Factor E. The following significant loadings appear in Factor E: 9. Eye-hand coordination, .91; 7. Dexterity of hands and arms, .78; 11. Coordination of independent movements of both hands, .78; 6. Dexterity of fingers, .60; and 1. Work rapidly for long periods, .30.

This factor appears to be primarily Manual Dexterity. Eye-hand coordination has the highest loading on this factor, but the distinction between this characteristic and Characteristic 7 is not entirely clear. The correlation between the two characteristics is .81. The example cited in the manual on Worker Characteristics for an A amount of Characteristic 7 is Drill-Press Operator (3, p. 16); and for Characteristic 9 the example cited is Engraver, Hand IV (3, p. 19). A case might be made for the interchangeability of the two examples. Whether the two characteristics are substantially the same is not, of course, established by this study. The fact of the relationship as here measured is merely noted.

Factor F. Factor F has the following significant loadings: 2. Strength of hands, .52; 42. Estimate quality of objects, .48; 3. Strength of arms, .40; 21. "Muscular" discrimination, .37; and 32. Understanding of mechanical devices, .34.

It is not believed that this factor is psychologically meaningful as it stands. Had enough care been taken it is possible that the factor would have converged with the A and C planes. Perhaps Characteristics 2 and 3 or Characteristics 21 and 42 might have been established as independent factors.

Factor G. Factor G has the following significant loadings: 46. Tools used, .55; 47. Knowledge of graphic instructions required, .52; 32. Understanding of mechanical devices, .43; 1. Work rapidly for long periods, −.40; and 21. "Muscular" discrimination, .39.

Factor G appears to be a Mechanical Information factor. The significance of the large negative loading on Characteristic 1 can only be surmised. Most psychological traits, such as intelligence, numerical ability, verbal ability, exist in positive or zero amounts, but not in negative amounts. However, negative traits are recognized. Thurstone notes, for instance, such opposite traits as "tactfulness" and "tactlessness" (13,

p. 165). In this instance, it is not known whether what is indicated is that workers with a fund of mechanical information do not work rapidly or that they do not work for long periods of time at a stretch.

Factor H. Factor H is a residual factor, with no significant loadings and apparently without psychological meaning: 16. Keenness of vision, .39; 21. "Muscular" discrimination, .30; and 27. Intelligence, .30.

Summary

The 20 Worker Characteristics, together with Skill and two other job characteristics, have been reduced to six meaningful factors: Strength, Intelligence, Inspection, Physically Unpleasant Working Conditions, Manual Dexterity, and Mechanical Information. This economy has been effected at the cost of a certain loss in specificity. Whether factor scores for the 275 occupations would be as valuable as a larger number of more specific scores depends on the use to which the information is to be put; just as the information contained in the Worker Characteristics Form is more and less valuable in various respects than the extended information in a voluminous job analysis. Certainly for the purpose of establishing a limited number (less than fifty) of occupational fields distinguished on the basis of worker characteristics for use in counseling, six factors are perhaps as many as can be considered. In this event it becomes important to find six independent and fundamental factors.

Such factors may be sufficient for the broad mass of industrial jobs. At the professional level they would not; there it would certainly be important to break up Intelligence at least into the verbal, numerical, and spatial factors commonly found in the literature.

A not inconsequential finding is that of the Strength factor. Aptitude tests for occupational selection do not ordinarily include a strength test, nor is it important that they should. But for many jobs at least a coarse evaluation of the strength of the applicant should be made in conjunction with and prior to test administration. There is little point in testing the manual dexterity of a physically weak applicant for a job which requires both strength and manual dexterity.

References

1. Chesire, L., Saffir, M., and Thurstone, L. L. *Computing diagrams for the tetrachoric correlation coefficient.* Chicago: University of Chicago Bookstore, 1933.
2. Coombs, C. H. A criterion for significant common factor variance. *Psychometrika,* 1941, 6, 267–272.
3. Employment Office Service Division. *Rating of worker characteristics.* Washington: War Manpower Commission, 1943. Multilithed.
4. Kelley, T. L. *Statistical method.* New York: Macmillan, 1924.

5. Landahl, H. D. Centroid orthogonal transformations. *Psychometrika,* 1938, **3,** 219–223.

6. McNemar, Q. On the number of factors. *Psychometrika,* 1942, 7, 9–18.

7. Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.

8. Shartle, C. L., Dvorak, B. J., and Associates. Occupational analysis activities in the War Manpower Commission. *Psychol. Bull.,* 1943, 40, 701–713.

9. Stead, W. H., and Masincup, W. E. *The occupational research program of the United States Employment Service.* Chicago: Public Administration Service: 1941.

10. Stead, W. H., and Shartle, C. L. *Occupational counseling techniques.* New York: American Book Company, 1940.

11. Thurstone, L. L. A new rotational method in factor analysis. *Psychometrika,* 1938, 3, 199–218.

12. Thurstone, L. L. *Primary mental abilities.* Chicago: University of Chicago Press, 1938.

13. Thurstone, L. L. *The vectors of mind.* Chicago: University of Chicago Press, 1935.

14. U. S. Employment Service, U. S. Department of Labor. *Dictionary of occupational titles.* Washington: U. S. Government Printing Office, 1939.

15. Viteles, M. S. Job specifications and diagnostic tests of job competency designed for the auditing division of a street railway company. *Psychol. Clinic,* 1922, 14, 83–105.

16. Viteles, M. S. *Industrial psychology.* New York: Norton, 1932.

Reported and Demonstrated Values of Vocational Counseling

Rose G. Anderson

The Psychological Corporation, New York City

The question most frequently asked by individuals considering vocational counseling is, "How successful is such counseling?" Consultants also have a vested interest in the answers to the questions, "What are the values which the individual feels he has gained from the counseling process?"; "What are the demonstrable practical outcomes of such counseling?"

There is a dearth of scientific evidence bearing on these questions. Much of the content of popular articles on the subject tends to mislead rather than inform the public as to the aims of qualified counselors and as to the justifiable expectations of the individual counseled.

Careful investigation in these areas is important because of the increased public interest and professional activity in vocational counseling. It is also essential for the contributions which may be made to improvements in counseling procedures.

Recognition of the desirability for such research is reflected in investigations which have been initiated and reported by Veterans' Advisement Centers (1, 2, 3).

Two independent but related counseling projects afforded the writer the opportunity to investigate: (1) the evaluations of their counseling experience by a group of civilian industrial employees; and (2) the practical outcomes of counseling for a group of ex-Service men.

The first of these projects was initiated by the Woodward Governor Company during World War II in the interests of the post-war readjustments of temporary employees and in the interests of the most effective use of the employees' skills. Comprehensive vocational counseling by a unit [1] directed by the writer was made available to all employees at company expense. A total of 1184 (85 per cent) of plant personnel at all occupational and executive levels availed themselves of this opportunity.

The counseling was completed the week of V-J Day. Prior to this period, a follow-up questionnaire designed to evaluate the benefits of the counseling experience had been circulated to 1086 counselees, 671 men and 415 women. The number of returns was limited by the dropping

[1] The full-time counselors were: Olive Bray, Ralph Filburn, and William Van Newkirk. Miss Bray supervised the tabulation and analyses of the questionnaire returns

of 600 employees the Monday after V-J Day, and by the fact that a considerable number of employees who had previously left the plant could not be reached. In addition, 51 men who had been counseled just before their induction into the Services were not sent questionnaires.

A total of 685 returns was received out of 1086 questionnaires distributed, a return of 62.1 per cent for men, 64.6 per cent for women, 63.1 per cent for the total group.

Excerpts are quoted from the statement accompanying the questionnaire: "The only interest in the returns is in getting a frank unbiased expression of opinion from each member as to what the counseling has meant to him or her. The counselors are not asking for bouquets, although they don't object to them. They are looking for any clews they can get as to whether, and if so how, the service can be made more helpful." The following statement was included from the company president: ". . . At the present time, Rockford industrial, commercial and agricultural leaders are seriously considering a vocational counseling set-up for returning service men from this area. You will help these men make a more intelligent decision in regard to such a program, and you will make a contribution to those who are developing and improving counseling procedures if you will give them the benefit of your experience." . . . "You will notice that a number appears in the space after *Name*. This is your code number known only to your counselors. You need not sign your name unless you wish to." . . . "The management will not read individual questionnaires and is not interested in *who* says *what*, but in the trend of results."

It would appear that the accompanying statements encouraged critical rather than complimentary answers.

Counseling Procedures

A preliminary description of counseling procedures is pertinent to the presentation of results.

Vocational counseling as practised in this project involved the integration of comprehensive measures of aptitudes and interests with: (1) detailed information about the individual's personal, educational, and vocational background; (2) the evaluation of personality characteristics and emotional adjustment through both interviews and objective measures; (3) available resources for training or avocational expression. The resulting integration was then used in helping the counselee to assimilate the interpretations evolving from his study, to arrive at decisions and implement them through appropriate measures.

A first step in setting up the project was a survey of community resources. This was ably administered by Dr. Ruth Cavan, a University of Chicago sociologist resident in Rockford. Dr. Cavan or her assistants personally interviewed the representatives of all institutions and agencies which offered educational, avocational, and vocational programs. All pertinent information

with respect to their offerings was compiled in a cross-referenced file for the use of the counselors. Information about educational resources outside of the city supplemented this file.

A second preliminary project was the assembling of a loan library in the plant. This included a comprehensive range of sources of vocational information, selected books on personal adjustment, child guidance, and adolescent psychology.

The plant project was initiated through orientation talks attended by all plant personnel. In these the procedures, aims and possible outcomes of counseling were discussed. Subsequent to this a common battery of tests was administered to all plant personnel in groups of 50–100. This served the dual purpose of introducing the testing program and of developing general and differential plant norms for the tests used. The preliminary battery of tests included the Adult Placement Test (a mental alertness test with verbal and numerical sections), the Psychological Corporation General Clerical Test, and the Revised Minnesota Paper Form Board Test. The Allport-Vernon Study of Values was used as an initial approach to interest trends. This measure was used with the expectation that it would be less applicable to those with limited educational backgrounds. Contrary to expectations, uniformly active interest was demonstrated by all.

After the completion of this testing schedule, each employee made his own decision as to whether he wished to continue with the comprehensive counseling. Individual interview appointments were then scheduled.

Prior to the individual interview, the counselor assembled the above test results, and the information in the plant personnel files. The latter included: (1) a comprehensive application questionnaire covering personal and family data, education and recreational activities; (2) previous employment references; (3) a plant employment record including periodic supervisors' evaluations and ratings, salary increases, transfers and promotions; (4) results of medical and visual examinations, and (5) the results of objective tests (chiefly trade information and arithmetic) administered as part of the application procedures.

In the first individual interview, the counselor corroborated and supplemented the above types of information, explored the counselee's attitudes toward his current position and job transfers, his educational objectives, and his social and personal relationships.

Two most important aspects of this interview were the establishing of a favorable relationship with the counselee and the opportunity to develop insights into his personality organization. At the termination of this first interview, the counselor asked the counselee to fill out the Bernreuter Personality Inventory, indicating it would be helpful in supplementing the interview. This was returned directly to the counselor. This practise was found to be essential if this instrument was to contribute any value.

On the basis of the interview and the compiled information, a supplementary appropriate testing battery was scheduled and individually administered by a psychometrist supervised by one of the counselors. The additional tests were chosen from a wide range of tests including intelligence tests, mechanical, artistic, musical, scientific, language, and clerical aptitude tests, achievement tests, and additional interest inventories.

A subsequent interview was scheduled when the counselor felt he had a basis for arriving at interpretations and recommendations. The integrity of the counselors in respecting the confidences of the employees led to counseling on intimate personal problems in many instances. Employees freely asked for further appointments. The plant physician cooperated actively on medical problems.

The maximum weekly case-load for each counselor was eight individuals. The same procedures were followed in working with the veterans with the

exception of greater dependence on the interviews for personal and background data.

Two of the counselors had no previous experience in vocational counseling, although both had experience in applied psychology. One held the M.A. degree in psychology, the other had completed his course requirements for the Ph.D. degree in psychology. The third counselor held a Master's degree in vocational education and had had extensive experience in vocational education, educational guidance, and in industry both as an employee and a consultant.

The first months of the project constituted an in-job-training program for the counselors under the writer's guidance through review of cases and staff conferences. Such guidance was continued in periodic visits throughout the period of both projects.

Name ....Date ....
Address ....Present Occupation ....
Draw a circle around the word which is your answer to each one of the questions on this page.

1. As a result of your counseling did you get a better idea
   a. Of your strongest abilities?........................ Yes No Doubtful
   b. Of your less strong abilities?...................... Yes No Doubtful
   c. Of your personality traits in relation to fields of work? Yes No Doubtful
   d. Of ways of promoting your personal development?.... Yes No Doubtful
   e. Of other fields of work or training you might transfer to? Yes No Doubtful
2. Will your counseling report influence your decisions as to future jobs or training?............................ Yes No Doubtful
3. Has your counseling report already influenced your decision as to a job or a field of training?..................... Yes No Doubtful
   If so, was the decision in accord with the results of your counseling?....................................... Yes No Doubtful
4. Did your test results show you had under-estimated your aptitudes
   a. For particular jobs?............................. Yes No Doubtful
   b. In general?.................................... Yes No Doubtful
5. Did you have ambitions which your test results did not support
   a. For particular jobs?............................. Yes No Doubtful
   b. In general?.................................... Yes No Doubtful
6. Did your counseling
   a. Increase your general self-confidence?.............. Yes No Doubtful
   b. Decrease your general self-confidence?.............. Yes No Doubtful
7. Did your counseling on the whole give you a better understanding of yourself?................................ Yes No Doubtful
8. Do you feel that your counseling was a worthwhile experience?....................................... Yes No Doubtful
9. Do you recommend that Woodward Governor continue such counseling?.................................. Yes No Doubtful
10. Would you recommend it to others at their own expense? Yes No Doubtful

Further Comment: (In the space below answer the question and make any further comments you would like to make. You may use the reverse side of the paper if you wish.)
At what age do you think such counseling should be provided?   ....years.

Fig. 1. Questionnaire used in counseling follow-up study.

General Results

The complete questionnaire is reproduced in Figure 1.

The analysis of the replies to the different items for men and women separately and combined, and for the plant employees and those who had left the plant are presented in Table 1.

Table 1

Questionnaire Returns from Individuals Counseled

No. Cases	Men in Plant 366 % Yes	Left Plant 51 % Yes	Women in Plant 171 % Yes	Left Plant 97 % Yes	Total 685 % Yes	% No	% Doubtful	% No Reply
Question								
1(a)	63	76	75	91	71	17	9	3
(b)	57	69	68	78	64	18	12	6
(c)	60	73	65	76	65	16	14	5
(d)	55	65	60	66	59	21	13	7
(e)	41	63	61	77	53	30	11	6
2	46	55	49	66	50	29	19	2
3(a)	29	47	18	43	29	61	8	3
(b)	24	43	17	36	26	18	5	51
4(a)	31	51	37	58	38	46	9	7
(b)	26	27	42	36	32	49	12	8
5(a)	33	27	29	40	33	54	6	7
(b)	18	18	24	26	20	60	7	13
6(a)	55	80	57	57	58	28	11	3
(b)	7	6	6	3	6	63	8	22
7	66	80	70	88	71	17	11	1
8	76	90	84	96	82	8	9	2
9	64	82	70	89	70	14	12	3
10	63	71	64	80	66	16	15	3

The positive responses for the total group of 685 are presented at the right side of the Table. The results indicate that: 71 per cent got a better idea of their strongest abilities; 38 per cent found they had underestimated their aptitudes for particular jobs, 32 per cent in general;[2] 71 per cent got a better understanding of themselves; 65 per cent and 59 per cent respectively got a better understanding of their personalities in relation to fields of work or of ways of promoting their personality development. Although 33 per cent reported ambitions not supported by test results for particular jobs, and 20 per cent reported general am-

[2] The amount of over-lap in these per cents is not available.

bitions not supported by test results, 58 per cent reported increased self-confidence; only 6 per cent, decreased self-confidence.[3]

With respect to the vocational questions, 53 per cent got a better idea of vocational transfer possibilities; 50 per cent expect future vocational or training decisions to be influenced by their counseling; 29 per cent had already made such decisions, 26 per cent in accord with the results of their counseling. With respect to the last two figures, it should be noted that 21.6 per cent of the total group had left the company and had had the occasion to make use of the counseling. The 51 per cent who did not reply to question 3(b) includes those who answered 3(a) in the negative. Some individuals answered "No" to both 3(a) and 3(b). This accounts for the apparent discrepancy between the "Yes" answers to 3(a) and the "Yes" and "No" answers to 3(b).

The lower per cents of positive replies to Questions 9 and 10 (70 and 66 per cent respectively) may suggest skepticism as to the sincerity of the replies to Question 8 (82 per cent). The ensuing discussion has a bearing on this.

The project was initiated by orientation discussions attended by all plant personnel. At this time, emphasis was placed upon the fact that the counseling was a cooperative affair and that individuals would profit most who had a genuine desire for better self-understanding. The suggestion was made that individuals should elect the counseling because of such a desire and not because of curiosity nor because the counseling was to be at the company's expense. A considerable number who did not request counseling at the beginning of the project later asked for it, after reports began circulating through the plant as to the results. The final number counseled included a certain proportion of employees who had no serious need for counseling nor intention to put the results to practical use. When this is taken into consideration, the indication is that the per cents for the positive replies are lower than they would be for individuals requesting counseling because of a felt need. The per cents answering Questions 9 and 10 affirmatively may represent the proportion of the total group who had the more serious interest in self-understanding.

Differential Results

Examination of Table 1 reveals generally more favorable replies by those employees who had left the plant than by those still employed; also more favorable replies by the women than by the men.

Ninety per cent of the men and 96 per cent of the women who had left the plant answered Question 8 in the affirmative; 71 per cent and 80 per

[3] The 22 per cent not replying to 6(b) include the 19 per cent who did answer 6(a).

cent respectively answered Question 10 in the affirmative. The majority of the employed women were war-time employees. The more favorable replies by those who had had the occasion to apply their counseling or who anticipated a change in employment are considered significant.

The responses reflecting positive contributions were analyzed according to age, years of education, and percentile on the Adult Placement Test (a mental alertness test with verbal and numerical sections) for both men and women.

The results for the men are reported in Table 2.

Table 2

Affirmative Responses of 366 Men to Counseling Questionnaire Classified by Age, Education and Tested Ability

	Age			Years—Education				Ability—Percentile		
	(18–25)	(26–40)	(41+)	(1–8)	(9–11)	(12)	(13+)	(1–30)	(31–70)	(70+)
No. Cases	28	254	84	57	70	166	73	91	134	141
	%	%	%	%	%	%	%	%	%	%
Question										
1(a)	68	67	50	53	73	64	58	60	61	66
1(c)	54	65	48	53	69	59	60	52	59	67
1(d)	46	59	45	46	59	58	53	53	53	59
1(e)	46	43	35	37	47	40	42	46	40	40
2	50	50	33	33	56	45	49	40	46	50
6(a)	57	55	56	54	64	54	51	47	59	57
7	64	67	61	63	69	65	66	64	64	68
8	79	75	76	67	76	77	79	69	77	79
10	54	64	63	65	67	57	71	66	60	63

According to expectation, the oldest men derive less personal benefit than the younger men. In spite of this, a third of the group indicate their future decisions will be influenced by their counseling. They report increased self-confidence in the same degree as the younger men. Also, they support self-financed counseling as strongly as the younger group.

In the analysis for years of education, the group with a grammar school education or less report less favorably than the other groups. However, 33 per cent report that future decisions will be influenced, 54 per cent report increased self-confidence, 67 per cent feel it was a worthwhile experience and 65 per cent recommend self-financed counseling. The group with (9–11) years of education report the most positive benefits. Favorable test comparisons for many in this group with the high school graduates contributed to the more positive replies. The results for the group who are high school graduates and the group with

some college are in general comparable. The latter group support self-financing more positively than any of the other groups.

In the case of the ability groupings, the highest ability group report most favorably. However, the uniformity in the ability groups is more striking than are the differences.

The counselors had anticipated less favorable replies for the men with fewer years of education and lower test results. In many cases, the limited educational background and low test results on all tests used necessitated basing recommendations chiefly on work experience and interests. Some counselees resented this, as indicated by the adverse comment by a "minishinest," aged 47, with less than an 8th grade education and a test percentile below 30: "The test only gave me the same thing as the concolr talked of so the time it took to answer was wasted. I think as they did not tell me anything I did not know. I expected to learn a lot but was dissiponted."

In contrast, another counselee aged 42 years in the same educational and test groups replied affirmatively to 1(a) (b) (c) (d), 5(a), 6(a), 7, 8, 9, and 10. He added the comment: "Let me add my thanks and appreciation to the parties who conducted the tests and interviews. They were handled in a very friendly manner that made one feel they had your personal interest at heart." The tabulated results reflect more support for the attitude reflected in the latter statement.

Table 3
Group Comparisons with Total Plant Men

| | Agree Total Plant Men | Disagree Total Plant Men | |
		Lower	Higher
Group I			
No. = 23	1(e)	1(a)	9
Age 41 or older	6(a)*	1(c)	
Median age 50	8	1(d)	
Education	10	2	
8th or less		3	
Percentile			
30 or less			
Group II			
No. = 32	1(a)	3(a)	1(e)
Age 26–40	1(d)		2
Median age 30	1(e)		10
Education	6(a)		
Some college	7		
Percentile	8		
70 or more	9		

* Only one of this group answered 6(b) in the affirmative.

A comparison is presented in Table 3 of the replies for men over 41 years in age with an 8th grade education or less, who fell below the 30 percentile on the Adult Placement Test, with the men aged 26 to 40 years with some college education, whose test percentiles were 70 or higher. Items on which each group agree or disagree with the total plant men are shown.

The older, less able group report less favorably than the total plant men on questions related to self-appraisal and vocational decisions; equally favorably on increased self-confidence, counseling as a worthwhile experience, and self-financing; and more favorably on counseling as a plant practice.

The younger, more able, better educated group report more favorably than the total plant men on possible job transfers, influence of counseling on future vocational decisions, and self-financing of counseling.

Table 4

Affirmative Responses of 171 Women to Counseling Questionnaire Classified by Age, Education and Tested Ability

	Age			Years—Education				Ability—Percentile		
	(18–25)	(26–40)	(41+)	(1–8)	(9–11)	(12)	(13+)	(1–30)	(31–70)	(70+)
No. Cases	81	55	35	13	31	89	38	33	73	65
	%	%	%	%	%	%	%	%	%	%
Question										
1(a)	80	65	80	54	71	78	81	67	77	78
1(c)	64	71	57	62	48	65	78	58	68	65
1(d)	64	62	46	46	55	62	63	52	66	57
1(e)	70	56	49	46	52	65	66	64	58	65
2	54	42	46	38	42	48	58	45	45	54
6(a)	60	49	60	54	58	53	66	55	52	63
7	74	58	77	69	61	71	74	76	67	69
8	86	84	80	69	84	82	95	70	85	91
10	67	65	57	46	55	66	74	48	66	71

In Table 4, the affirmative responses for the women are presented. Positive values are reported for all three age and educational levels, with the older, less well educated women reporting less favorably on ways of promoting personality development and job transfers.

Positive values are reported by all ability levels with the least favorable replies for those in the lowest ability level for questions bearing on personality in relation to fields of work, promotion of personality development and self-financing of counseling. Both the middle and low ability groups report less favorably on effect of counseling on future vocational

decisions. In general, the most favorable responses are for the highest ability group.

The majority of the older women were housewives working in the cafeteria or at unskilled jobs who were not contemplating post-war employment.

As previously indicated, this survey was conducted for the enlightenment of the counseling staff with respect to positive values reported and with respect to possible improvements in the procedures.

A higher proportion of returns is considered desirable for representative conclusions. Earlier in the project, however, the president of the plant personally conducted his own survey of employees who had been counseled. At that time, he received 80 per cent of favorable replies. The counseling unit felt that the group which had been counseled to that point was rather heavily weighted with older or less able men. This latter fact and the agreement of the earlier survey with the questionnaire returns provide some basis for regarding the latter as representative.

Attitudes Reflected in Counselees' Comments

Popular, much publicized articles have over-emphasized "the hidden talents" and "the many aptitudes" revealed by aptitude testing. Consequently, many individuals whose results do not reveal startling "new directions" or a "highway to success" experience a natural disappointment.

Qualified vocational counselors are modest in their claims as to the benefits of counseling. They are aware of the limitations in their techniques, of the many economic and social factors preventing full use of recommendations, and of the emotional resistances in counselees against accepting and acting upon interpretations which conflict with their own self-evaluations or their needs for emotional security and prestige.

A considerable proportion of respondents added comments to supplement their replies. Samples which reflect various attitudes of both a critical and a favorable nature merit report and discussion.[4]

Skepticism is expressed by a young college man in the high ability group as to whether the counselee should be given his results: "It is O.K. for the employer, but how much the individual should or need be told concerning his merits is debatable. Many middle group individuals are better performers than the so-called top-flight brains, because of the differences in character which tip the scales back to offset the original advantage. Why discourage the plugger by telling him he can't possibly reach his goal? Many will if they don't know that they haven't got the capacity." The evaluation of those compensatory assets which offset

---

[4] Other comments were included in the Psychological Service Center Bulletin for April, 1946.

A comparison is presented in Table 3 of the replies for men over 41 years in age with an 8th grade education or less, who fell below the 30 percentile on the Adult Placement Test, with the men aged 26 to 40 years with some college education, whose test percentiles were 70 or higher. Items on which each group agree or disagree with the total plant men are shown.

The older, less able group report less favorably than the total plant men on questions related to self-appraisal and vocational decisions; equally favorably on increased self-confidence, counseling as a worthwhile experience, and self-financing; and more favorably on counseling as a plant practice.

The younger, more able, better educated group report more favorably than the total plant men on possible job transfers, influence of counseling on future vocational decisions, and self-financing of counseling.

Table 4

Affirmative Responses of 171 Women to Counseling Questionnaire Classified by Age, Education and Tested Ability

	Age			Years—Education				Ability—Percentile		
	(18–25)	(26–40)	(41+)	(1–8)	(9–11)	(12)	(13+)	(1–30)	(31–70)	(70+)
No. Cases	81	55	35	13	31	89	38	33	73	65
	%	%	%	%	%	%	%	%	%	%
Question										
1(a)	80	65	80	54	71	78	81	67	77	78
1(c)	64	71	57	62	48	65	78	58	68	65
1(d)	64	62	46	46	55	62	63	52	66	57
1(e)	70	56	49	46	52	65	66	64	58	65
2	54	42	46	38	42	48	58	45	45	54
6(a)	60	49	60	54	58	53	66	55	52	63
7	74	58	77	69	61	71	74	76	67	69
8	86	84	80	69	84	82	95	70	85	91
10	67	65	57	46	55	66	74	48	66	71

In Table 4, the affirmative responses for the women are presented. Positive values are reported for all three age and educational levels, with the older, less well educated women reporting less favorably on ways of promoting personality development and job transfers.

Positive values are reported by all ability levels with the least favorable replies for those in the lowest ability level for questions bearing on personality in relation to fields of work, promotion of personality development and self-financing of counseling. Both the middle and low ability groups report less favorably on effect of counseling on future vocational

decisions. In general, the most favorable responses are for the highest ability group.

The majority of the older women were housewives working in the cafeteria or at unskilled jobs who were not contemplating post-war employment.

As previously indicated, this survey was conducted for the enlightenment of the counseling staff with respect to positive values reported and with respect to possible improvements in the procedures.

A higher proportion of returns is considered desirable for representative conclusions. Earlier in the project, however, the president of the plant personally conducted his own survey of employees who had been counseled. At that time, he received 80 per cent of favorable replies. The counseling unit felt that the group which had been counseled to that point was rather heavily weighted with older or less able men. This latter fact and the agreement of the earlier survey with the questionnaire returns provide some basis for regarding the latter as representative.

Attitudes Reflected in Counselees' Comments

Popular, much publicized articles have over-emphasized "the hidden talents" and "the many aptitudes" revealed by aptitude testing. Consequently, many individuals whose results do not reveal startling "new directions" or a "highway to success" experience a natural disappointment.

Qualified vocational counselors are modest in their claims as to the benefits of counseling. They are aware of the limitations in their techniques, of the many economic and social factors preventing full use of recommendations, and of the emotional resistances in counselees against accepting and acting upon interpretations which conflict with their own self-evaluations or their needs for emotional security and prestige.

A considerable proportion of respondents added comments to supplement their replies. Samples which reflect various attitudes of both a critical and a favorable nature merit report and discussion.[4]

Skepticism is expressed by a young college man in the high ability group as to whether the counselee should be given his results: "It is O.K. for the employer, but how much the individual should or need be told concerning his merits is debatable. Many middle group individuals are better performers than the so-called top-flight brains, because of the differences in character which tip the scales back to offset the original advantage. Why discourage the plugger by telling him he can't possibly reach his goal? Many will if they don't know that they haven't got the capacity." The evaluation of those compensatory assets which offset

---

[4] Other comments were included in the Psychological Service Center Bulletin for April, 1946.

less favorable test results constitutes the chief difference between comprehensive vocational counseling and aptitude testing. The respondent is not speaking from first-hand experience when he implies that the counselor tells the individual "he can't possibly reach his goal." The emphasis in each case is on the most positive potential possibilities of the individual. In arriving at these, some goals are indicated as less promising in returns for effort expended.

Another common misunderstanding is reflected in the comment of a fifty year old man who was a war-time employee. He had formerly been a photo-engraver. His results provided most positive support for returning to his former occupation. He stated, "I don't want to because I know for years that I wasn't a whiz at that job and never would be. . . . I'd say that for me it (the counseling) was a bit wasteful of time although I found it interesting." A point requiring constant re-interpretation is the fact that recommendations for certain kinds of work do not carry an inherent guarantee of successful competition with others so engaged. On the contrary they indicate that chances for productivity are relatively better in the suggested fields than in other fields considered. From this man's comment, it is apparent that this was not sufficiently clarified in his interviews with the counselor.

Another comments, "I enjoyed taking the tests and talks with the counselor. . . . What I objected to most was methods of comparison. Instead of using John Doe, same age, same type community and same general type, we were given dozens of comparative types." This point is well taken. However, comparisons must necessarily be made in terms of standardization groups. This criticism overlooks the fact that plant norms were established for three of the basic tests given to the entire plant personnel. Norms were established for 16 occupational groups, 10 educational levels, and for the total plant group for the Adult Placement Test and the Psychological Corporation General Clerical Test. Norms for four occupational and two age groups for men, and for two occupational groups and two age groups for the women were established for the Revised Minnesota Paper Form Board Test. Comparisons were reported with the appropriate plant groups as well as with the standardization groups. The "dozens of comparative types" mentioned suggests the possibility that this counselee was given more comparisons than he needed or could assimilate.

Another man reports, "I sincerely believe that I wasted my time and the company's taking the tests. I was not in a proper frame of mind when I took them. . . . I would sincerely like to go through it again with ———— instead of ————." There was evidence that the personal factor of the counselor did influence the attitudes toward the counseling

results. An analysis of the questionnaires according to who had done the counseling reflected more favorable replies for certain of the counselors.

In many instances, transfers were made within the plant on the basis of the counseling results. Apropos of such changes is the following comment: "Since the counseling I have been shifted by the management into a job that is much more in line with my aspirations. I am very satisfied with the change."

Excerpts from other favorable comments include: "I think it was a great opportunity and intend to follow it as closely as possible." "I think it is a good thing because it is so easy to be mistaken by wishful thinking." ". . . It helped me a great deal in planning my future." "I used my counselor's summary to good advantage in obtaining the above position." ". . . Gave me a better idea of opportunities for employment of one of my age." ". . . Gave me confidence in myself to do the kind of work I've always wanted to do." ". . . Gave me good insight into the future in the respect of family affairs." ". . . Received several helpful suggestions that have proved beneficial in my everyday life."

Placement Follow-up of Veterans

The second project referred to above afforded the opportunity to check placement outcomes against counseling recommendations. The more subjective evaluations of the benefits of counseling could be compared with practical results.

A Veterans' Information and Placement Service was set up and financed by Rockford business men and industrialists just subsequent to the completion of the plant project. The vocational counseling for this project was handled by the same counseling unit which functioned in the Woodward-Governor Company.[5] A full time placement officer worked independently but cooperatively with the counseling unit in finding employment for the veterans.

The veterans comprise the age-range least represented in the plant studies. As a group, they represent a higher educational level than the plant men. Fifteen per cent had less than a high school education, compared to 34.7 per cent for the plant; 61 per cent were high school graduates, compared to 45.3 per cent for the plant; and 24 per cent had had some college training or were college graduates, compared to 20 per cent for the plant. Twenty-two per cent were commissioned officers, 78 per cent were non-commissioned officers, technicians, or privates in the Services.

[5] Miss Olive Bray left the staff at this time to set up a counseling unit at Rockford College for Women.

When he applied for the counseling service, the veteran agreed that the placement officer was to be given a copy of the vocational recommendations. In order to make the latter as meaningful as possible, a code was developed which indicated whether the veteran had the supporting qualifications for a suggested occupation in high degree, in moderate degree, or in questionable degree; also whether he needed further education, on-the-job training, or work experience.

Since a number of alternatives was listed for each veteran, the code enabled the placement officer to guide the men into those activities for which they were best qualified. Periodic check-ups were made with both the employee and the employer after placement. The records of placement and follow-up were kept by the placement officer. The analysis of the results was made by him and made available to the writer.

The results reported were compiled at the time that 516 veterans had completed their counseling. Of this number, 444 were available for follow-up in Rockford. Of these, 82.4 per cent were satisfactorily placed in recommended jobs according to their own and their employers' statements; 10.9 per cent had been placed in other jobs. At the time of the last follow-up, 7 per cent were not yet employed. The last group included a number of men in upper economic brackets who had not sought employment.

The employment stability record was one of the most outstanding results of the veterans' placement service. Of those employed, only eleven men, less than 3 per cent, had changed jobs in a period of 19 months.

Labor turn-over was high in the early post-war period. For the mid-months of the veterans' counseling project (January and February, 1946), the monthly turn-over for eleven major U. S. industries (5) ranged from a low of 5.3 per cent to a high of 13.1 per cent; for manufacturing it was 6.8 per cent. Since Rockford is a manufacturing center, the comparison between the latter figure and the record of the veterans' placement service provides some basis for judging the value of counseling prior to placement.

Summary

Analyses of replies to a questionnaire designed to gauge counselees' estimates of the benefits of counseling have been presented. The results were analyzed separately for employees remaining in their jobs and those who had left for other work or training. Results were also analyzed according to age, education, ability level, and sex of counselees.

The trend of the questionnaire returns provides strong confirmation that the counseling had significant positive values for the participants. The results also confirm the philosophy underlying the framing of the

questionnaire. That is, that the values vary from individual to individual. The fact that the highest per cent (82) of positive replies is to Question 8, "Do you feel that your counseling was a worthwhile experience?", may be regarded as evidence that the values are distributed among the areas covered by the other questions. More favorable reports were made by those who had had the occasion to use the results of their counseling, and by those who anticipated post-war job transfers.

Volunteered comments have been reported which throw light upon some of the common misconceptions of counseling, on necessary cautions to be observed by counselors, and on the varied positive contributions to the individuals counseled.

A supplementary report is included of placement follow-ups for veterans who had been placed in accordance with the results of their counseling. Of 444 veterans available for follow-up, 82.4 per cent were satisfactorily placed according to their own and their employers' statements. During a period of 19 months, less than 3 per cent had changed jobs.

This report has not included the replies to the final question on the questionnaire with respect to recommended age for counseling. These results have been analyzed in a separate report (4).

References

1. Long, L., and Hill, J. A follow-up study of veterans receiving vocational advisement. *J. consult. Psychol.*, 1947, 11, 88–92.
2. Brown, M. T. The veterans report one year later. *Occupations*, 1947, 25, 209–212.
3. McCue, E. F. Summary report of veterans in training. *Occupations*, 1947, 25, 340–342.
4. Anderson, Rose G. Preferred ages for vocational counseling. *Occupations*, 1948, 27, 77–81.
5. Monthly Labor Review, U. S. Bureau of Labor Statistics, May, 1946.

Vocational Interests of Accountants

Edward K. Strong, Jr.

Stanford University

In scoring Strong's Vocational Interest Blank, three scales have been employed to measure the interests of men employed in office-accounting work (5). These are:

1. Office worker, representative of office activities in business concerns including bookkeepers, purchasing agents, credit managers, and office managers.

2. Accountant, based upon the records of 160 general accountants, 54 cost accountants, 65 auditors, and 66 comptrollers and treasurers.

3. Certified Public Accountant, so certified in the states of New York and California.

Accountants and CPA's were further differentiated in developing the scales so that the former represented men regularly employed by business firms whereas the latter were employed by public accounting concerns. Among the former were some men holding the CPA certificate. But they were classified as accountants, not CPA's on the basis of the employer for whom they worked rather than on the basis of whether or not they had the CPA certificate.

A fourth scale is related to the above, i.e., Purchasing Agent, composed entirely of purchasing agents, whereas the Office Worker Scale is primarily representative of bookkeepers and related office activities with a smaller representation of men holding more advanced positions in office work.

Since these scales were developed there have been many queries regarding the relationship between the Accountant and CPA Scales, because each group scored about 40 on the other scale but the correlation between them was only .28.

The primary question concerning us here is: do the Accountant and CPA Scales measure what they purport to measure?

In 1943 a survey was made of members of the American Institute of Accountants by Dr. Ben Wood and Dr. A. E. Traxler, under the direction of the Committee on Selection of that Institute. The survey included tests of: (a) ability or achievement, with which we are not here concerned; and (b) the Vocational Interest Test. Data on the latter were obtained from 1856 accountants (1). Additional records of 1117

accountants in Canada were secured under the auspices of the Dominion Association of Certified Accountants (2).

Occupational interest scores were determined for 1000 of the American accountants, 200 from each of five sub-groups, namely, partners, managers, seniors, semi-seniors, and juniors. The median score on each of twenty-three scales agreed very closely with the medians based on the 1117 Canadians.

On the basis of these data the present Accountant Scale is judged adequate for juniors since the distribution of letter ratings of 314 juniors agreed very closely with the distribution of ratings of the criterion group. See Table 1.

Table 1

Percentage Distribution of Letter Ratings on Accountant Scale of 314 Junior Accountants and Members of Criterion Group upon which the Scale is Based (2)

Letter Rating	Juniors	Criterion Group
	%	%
A	65.0	69.2
B+	16.2	15.0
B	12.1	9.2
B—	3.2	4.4
C+	2.9	1.6
C	0.6	0.6

At first thought it is surprising that the Accountant Scale should represent the interests of junior accountants. The Accountant Scale, as pointed out above, represents not merely men of junior level but also to some extent men of higher levels up to the top ranks of comptroller and treasurer. The explanation may lie in the likelihood that the group of juniors contains within it men who are destined, later on, to reach intermediate and top levels of accounting work in a business concern. If this is the situation it is support for the writer's procedure of developing scales for an occupation rather than a position or level within an occupation.

The present CPA Scale, on the other hand, does not properly represent the interests of the great majority of public accountants. See Table 2. None of the sub-groups, except partners, scores high enough for the Scale to be considered as representative of their interests. The distribution of scores of partners is such that the Scale can be used to represent their interests fairly well, although it is to be expected that partners will secure somewhat fewer A ratings and more B and B— ratings than should be expected from a scale that adequately represents them.

This raises the question, when may two groups be considered to be one group and when two separate groups? What objective criteria may be set up to answer this question? Use of critical ratios, or their equivalent in terms of level of significance, is not applicable. Two means may be significantly different and at the same time common sense indicates that the two groups differ too little to be divided into two separate groups. Percentage of overlapping appeals to the writer as probably the best criterion to use in this connection.[1] There is, however, no

Table 2

Percentage Distribution of Letter Ratings on Original CPA Scale of Five Levels of Accountants and of Members of the Criterion Group (2)

Letter Rating	Criterion Group	283 Partners	226 Mgrs.	582 Snrs.	361 Semi-Sen.	311 Juniors	1766 Total
	%	%	%	%	%	%	%
A	72.6	60.8	44.3	36.3	38.8	31.2	41.0
B+	13.6	14.8	19.0	20.3	16.6	18.5	18.2
B	6.2	12.0	20.4	19.1	18.3	16.9	17.5
B−	3.7	8.5	9.7	12.7	12.7	17.2	12.5
C+	2.5	2.8	4.4	6.7	7.2	11.1	6.7
C	1.4	1.1	2.2	4.4	6.4	5.1	4.1

accepted agreement as to what percentage of total overlapping should be used as the cutting point. Using the data in Table 2, we have 87.9 per cent overlapping between scores of the criterion group and partners but only 71.7 per cent between the former and managers. Our present judgment is that 88 per cent is too high an overlapping to consider the two groups separate groups and that 72 per cent, on the other hand, is too low to include the two in one group. In making this statement we have in mind the overlapping of a considerable number of pairs of groups. One difficulty in arriving at the proper cutting point lies in the fact that popular opinion respecting whether two groups are similar or not and percentage of overlapping do not correlate perfectly.

Senior CPA Scale

A new scale has been developed based on the interests of 611 Senior CPA's. This Scale will be designated as "Senior CPA" in contradistinction to the old CPA Scale, to be referred to from now on as the Partner CPA Scale. The new Scale has a reliability of .89, which is much higher than the old Scale.

[1] Percentage of overlapping is the percentage of scores of one group which may be matched with scores in the other group, (8).

Table 3

Mean Scores of Office Men and Three Groups of Accountants on Their Four Scales and Also on the OL Scale

| | Occupational Groups | | | | | | | |
| | Office Man | | Accountant | | Senior CPA | | CPA Partner | |
Scales	M	σ	M	σ	M	σ	M	σ
Office Man	49.4	10.4	46.9	9.8	43.9	9.5	35.8	11.0
Accountant	43.6	11.4	50.3	9.6	47.7	8.4	41.1	11.7
Senior CPA	—	—	—	—	49.7	10.1	—	—
CPA Partner	31.6	9.8	39.3	11.0	42.4	9.4	50.1	10.4
OL	57.1	7.6	59.6	8.0	57.4	6.3	62.7	7.0

The relationship of the four scales of Office Worker, Accountant, Senior CPA and Partner CPA are shown in Table 3. Scores on the Office Scale go down from office worker to CPA Partners and, in reverse, scores go up on the Partner CPA Scale from Office Worker to CPA Partner. Senior CPA's who hold an intermediate position between juniors and partners have interests more akin to juniors than to partners as far as scores on the four scales go. This is also true with respect to scores on the OL Scale.

Table 4

Correlation between Interests of Senior CPA's and Other Occupations and Mean Scores of Senior CPA's on the Various Scales

r	Scale	Mean	r	Scale	Mean
.72	Mathematics-Science Teacher	33.6	.10	Banker	35.7
.72	Accountant	47.7	.07	Partner CPA	42.4
.64	Policeman	30.2	.06	Mathematician	21.7
.58	Office Worker	43.9	.03	City Sch. Supt.	23.9
.53	Production Manager	38.6	.02	Psychologist	15.9
.53	Printer	30.8	−.01	Musician	23.6
.52	Aviator	29.3	−.07	Minister	16.4
.52	Forest Service	24.6	−.08	Dentist	20.8
.50	Carpenter	20.6	−.21	Architect	20.7
.47	Public Administrator	39.9	−.29	Physician	23.8
.41	Personnel Manager	35.3	−.36	President	32.6
.39	Y. Physical Director	25.8	−.40	Realtor	34.1
.38	Farmer	30.6	−.40	Life Insurance	28.7
.33	Purchasing Agent	38.7	−.55	Artist	16.7
.31	Engineer	31.8	−.55	Lawyer	30.2
.26	Chemist	27.8	−.61	Advertiser	28.4
.20	Social Science Teacher	29.1	−.70	Author	26.1
.19	Y.M.C.A. Secretary	24.3	−.64	OL	57.4
.16	Sales Manager	32.4	.59	MF	48.2

The correlation of interests of Senior CPA's and men in 36 occupations are given in Table 4. The correlations with accountant and office worker are .72 and .58, but only .07 with partner CPA's. The latter low correlation is similar to .28 between accountant and partner CPA and .06 between office worker and partner CPA. The interests of CPA partners are clearly quite distinct from the interests of office workers, junior and senior accountants. This relationship is shown clearly in terms of OL scores. Partner CPA correlates .43 with OL whereas the other three scales correlate negatively with OL, i.e., −.26 with accountant, −.33 with office man and −.64 with senior CPA's. The difference here is very large between partner CPA and the other three scales.

In line with the low negative correlation with OL the interests of senior CPA's are correlated positively with occupations in Group IV, i.e., mathematics-science teacher, policeman, printer, aviator, forest service, carpenter and somewhat lower with farmer. Actually the Senior CPA Scale correlates on the average of .54 with the seven occupations in Group IV, whereas the Accountant Scale correlates only .15, the Office Scale .09, and the Partner CPA Scale correlates −.44. This is a most unexpected relationship. The interests of Group IV typify mechanical interests. Why should senior accountants exhibit such interests and also why should they exhibit more mechanical interest than partners, on the one hand, and juniors on the other hand?

To what occupational group should the new scale be assigned? Up to the present time Group VIII has contained purchasing agent, office worker, accountant and banker. But senior CPA cannot be grouped with these four since it correlates only .33 with purchasing agent and .10 with banker. On the basis of correlations of .60 and over senior CPA's should be grouped with mathematics-science teacher (.72), accountant (.72), policeman (.64), and, if we stretch a point, office worker (.58). If high mean scores are considered, senior CPA should be grouped with accountant (score of 47.7 on that scale), office worker (43.9), partner CPA (42.4); and in the neighborhood of public administrator (39.9), purchasing agent (38.7), and production manager (38.6).

Ordinarily, as the correlation between two interests goes down, the mean scores of each on the other's scale also go down. But here we have an exception to this relationship. For example, the interests of senior CPA's correlate .72 with mathematics-science teacher but they average only 33.7 on that scale. Similarly the correlation with policeman is .64 but the mean score on that scale is 30.2. In grouping occupations it seems desirable to take both correlation and mean score into account. The writer, however, has found no statistical way of combining the two measures. About all that can be done is to group occupations on the

basis of both measures on a common sense basis. Upon this basis we would disregard the high correlations between senior CPA and both mathematics-science teacher and policeman because of the low mean scores. Another reason for doing this is that occupations having no obvious connection in everyday life should not be included in the same group. To ignore this criterion might result in occupational groups which are useless as far as guidance and selection are concerned (5).

For the time being, we suggest a division of the present Group VIII into Group VIIIa composed of purchasing agent and banker and Group VIIIb composed of accountant, senior CPA and office worker. The two sub-groups differ too much to be considered as one occupational group.

In terms of the Interest Global Chart, senior CPA should be located below accountant in the direction of Group IV but its exact location can only be determined by factor analysis.

CPA Partner Scale

We hope to revise the original CPA Scale before long, from now on to be called the CPA Partner Scale. Until that is done we believe the present scale is of distinct value. It reflects interests of a managerial type not particularly reflected in either the new Senior CPA Scale or the old Accountant and Office Worker Scales. The data in Table 2 make clear that scores on this scale decrease as one goes from partner to manager to rank-and-file accountant.

The fact that the CPA Partner Scale correlates highest of all with the Lawyer Scale (i.e., .57) but only .28 with accountant, .07 with Senior CPA and .06 with office worker, indicates to the writer that the CPA Partner and Lawyer Scales express the interests involved in dealing with a client regarding accounting or legal matters, an activity not common to the non-managerial employee.

Again and again in studying the interests of members of an organization the fact is brought out that there are real differences in the interests of the rank-and-file of employees and the interests of the administrators or executives in the top levels of management (4, 6, 7). We believe such differences are reflected in the noticeable differences in scores between the Senior CPA and the CPA Partner scales.

Existing data indicate that only a small minority of the rank-and-file of employees possess the interests that characterize the top managerial group. The evidence so far supports the assumption that it is from this minority that the future managers will come. If this is true it is most important to identify the minority early in their employment and to afford them special opportunities to prepare for advancement. The writer does not feel that the above point of view is more than a good

working hypothesis today. It is possible that interests may change with promotion and increasing responsibilities. If this proves to be correct it means that an organization should give attention to the early training of its superior subordinates not only in procedures but also as regards interests. How the latter is to be done, if it can be done, is certainly not definitely formulated today.

Occupational Interest Scores of Accountants

Table 5 gives the occupational interest scores of six groups of accountants on the 13 scales on which 1000 public accountants score the highest. The 1000 accountants are composed of 200 each of partners, managers, seniors, semi-seniors and juniors. The scores of students in accounting

Table 5

Occupational Interest Scores of Six Groups of Accountants

Note: Data in first three columns are medians, after Wood and Traxler; remaining data are mean scores.

	1000 Public Account- ants	100 Acct. Seniors	100 1st Year Students	100 Senior CPA's	100 Account- ants	100 Office Men
Accountant	47.5	44.6	40.3	47.7	50.3	43.6
CPA Manager	42.5	36.8	32.8	42.4	39.3	31.6
Office Man	—	—	—	43.9	46.9	49.4
Production Manager	39.8	38.2	37.0	38.6	39.2	36.2
Purchasing Agent	39.6	38.5	39.6	38.7	41.9	42.2
Banker	36.7	37.5	35.3	35.7	39.0	37.8
Personnel Manager	35.2	40.7	40.2	35.3	34.6	29.6
President	35.2	34.5	35.8	32.6	34.4	34.4
Realtor	34.7	—	—	34.1	35.6	39.2
Sales Manager	33.9	38.4	40.3	32.4	35.5	36.9
Math.-Science Teacher	33.4	33.2	31.2	33.7	30.1	27.9
Engineer	32.3	23.2	22.6	31.8	28.1	25.7
Lawyer	31.4	30.8	30.8	30.1	29.3	28.2
OL	—	—	—	63.1	59.6	57.1

approximate the scores of the thousand accountants but differ by having higher scores in personnel and sales management and lower scores on accountant, engineer, and especially CPA Partner. The differences are presumably due to the expectation that some of the students will not go into accounting work and that these students have less interest in accounting than the remainder.

The office men and accountants, representative of juniors, differ from

seniors in higher scores on office interest and lower scores on CPA Partner interest and OL.

Unfortunately Wood and Traxler did not score their blanks on the Office Men and OL Scales. These scales throw as much light on the relationship between the levels of accountants as any. The writer has found that the office man scale is the best single scale for indicating business interest. All business men from president to office man spend a good share of their time looking at and shuffling papers. When a counselee scores low on this scale one should have him consider other occupations than typical business activities.

References

1. *Report of Committee on Selection of Personnel.* American Institute of Accountants, Jan. 15, 1945.
2. Committee on Selection of Personnel. *A study of the ability of accounting students.* American Institute of Accounting, 1946, Bull. No. 1.
3. *The college and professional accounting testing programs.* American Institute of Accounting, 1947, Bull. No. 3.
4. Marsh, Commander James S. *Selection of civil engineer corps of Officers of United States Navy.* 1949, M.A. Thesis, Stanford University.
5. Strong, E. K., Jr. *Vocational interests of men and women.* Stanford University Press, 1943.
6. Strong, E. K., Jr. The interests of forest service men. *Educ. and Psychol. Measmt.,* 1945, 5, 151–171.
7. Strong, E. K., Jr. Interests of senior and junior public administrators. *J. appl. Psychol.,* 1946, 30, 55–71.
8. Tilton, J. W., *Measurement of overlapping. J. educ. Psychol.,* 1937, 28, 256–62.

Vocational Interests of Psychologists *

Philip H. Kriedt

Prudential Insurance Company, Newark, N. J.

This study was undertaken with the general objective of making the Strong Vocational Interest Blank a more useful tool for the guidance of both beginning and advanced psychology students. More specifically, the study proposed: 1. to determine the adequacy of the 1938 S. V. I. B. (Strong Vocational Interest Blank) psychologist key and to construct a new key if it seemed necessary; 2. to develop interest profiles for various sub-groups of psychologists based on the scores of these sub-groups on all the 1938 S. V. I. B. keys; and 3. to construct new keys for several sub-groups of psychologists so that better differentiation between sub-groups could be secured.

The Pilot Study

As a pilot study, 95 prominent psychologists who could be classified quite clearly as experimental, social, guidance, statistical, or industrial psychologists were asked to fill out the S. V. I. B. By using three follow-up letters, returns were received from 92 of the 95. Analysis of these data indicated that the 1938 psychologist key was not satisfactory for this group. Only 56% of them received scores of A or B+ instead of the 84% that would be secured if the key were completely appropriate. This difference is statistically significant (p. <.01). Analysis of the interest profiles of the five sub-groups showed sufficient differences between the sub-groups to warrant the collection of more data of this sort in order to construct sub-group keys.

The Main Study

For the main study, begun in April, 1948, all male psychologists who had received a Ph.D. before 1943 and whose addresses were listed in the 1948 APA Directory were asked to fill out a S. V. I. B. Three follow-up letters were needed in order to obtain 1048 (89%) usable returns. Analysis of the age and major field of experience of those who had not replied indicated that the sample obtained is not a biased one in those respects.

Each psychologist furnished information regarding his professional experience which made it possible to classify him according to "field" and in most cases according to "function" also. The numbers in each field classification are: 256 experimental, 221 clinical, 154 educational, 115 guidance, 108 industrial, 69 social, 65 statistical, 44 child, and 16 marketing. The numbers in each functional classification are: 295 teaching, 184 research, 146 service, and 128 administration. (There were 315 who were not given a functional classification.)

Profile Analysis

Median scores on the 42 present keys of the S. V. I. B. for these 13 sub-groups are presented in Tables 1 and 2.[1] Psychologists-in-general have median scores of A on the psychologist and public administrator keys; median scores of B+ on the chemist and personnel director keys; and median scores of B on the artist, architect, physician, mathematician, engineer, math and physical science teacher, city school superintendent, advertising man, lawyer, and author-journalist keys.

The rank order correlations between the median profile for all psychologists vs. each of the 13 sub-group profiles are as follows: clinical .98, social .96, child .94, educational .93, statistical .90, experimental .86, guidance .77, marketing .47, teaching .96, service .90, research .86, and administration .81. Marketing psychologists are the most deviant group and their profile shows that they are characterized by stronger sales and office detail interests than other psychologists. Guidance psychologists deviate in the social service direction. Research, experimental, and statistical psychologists are distinguished because of their physical science-biological science interests. Administrators have higher social service and production manager scores than most psychologists. The industrial sub-group differs from psychologists-in-general largely because of stronger production manager and office detail interests. The service sub-group have verbal and social service interests which distinguish them. The other sub-groups (teaching, child, educational, social, and clinical) have median profiles which are very similar to the total group profile.

The 1948 Psychologist Key

A new psychologist key was constructed by contrasting the responses of these 1048 psychologists with those of Strong's 1938 professional men-

---

in-general group. Strong's method of assigning item weights was followed. This new key differentiates psychologists from professional men-in-general rather sharply. All of the psychologists exceed the mean score for professional men-in-general. The mean standard score for professional men-in-general is 17.5 as compared to 50 for psychologists. In other words, using the standard deviation of the psychologist group as the unit of measurement, the two means are 3.25 standard deviations apart.

All 1048 psychologists were scored on this new key. Except for the marketing psychologists who have a mean score of B+, all the sub-groups have a mean score of A. Social, statistical, child, clinical, experimental, and research psychologists have high A mean scores; service and teaching psychologists have average A mean scores; guidance, educational, industrial, and administrative psychologists have low A mean scores.

All 13 sub-groups score higher on the 1948 psychologist key than they do on the 1938 psychologist key. Industrial, guidance, and administrative psychologists show the greatest increase in scores and experimental, teaching, and research psychologists show the least increase. This means that if an individual has the interests of one of the last three types of psychologists, the 1938 key will differentiate his interests from those of professional men-in-general quite accurately. If, however, he has the interests of an industrial, guidance, or administrative psychologist, the 1948 key is much more likely to give A and B+ ratings and presumably is a better reflection of the interest patterns of present-day psychologists than the 1938 key.

Examination of the item weights for the 1948 psychologist key, presented in Table 3, shows the ways in which the 1048 psychologists included in this study differ from Strong's 1938 criterion group.[2] The 1948 group is more socialized than Strong's group, more interested in, more tolerant of, and more willing to help people, and less interested in mechanical and methodical work and in solitary activity. The 1948 group seems to have more of the interests of an applied psychologist and Strong's group more of the interests of a laboratory scientist. The response weights for the two keys are different in 484 instances. Most of the differences are changes of only one point, but 41 shifts involve differences in weights of two or more points.

The practical significance of the difference between the two keys is indicated most clearly by the scatterplot of individual scores on the two keys, presented in Table 4. The data show that these psychologists tend to score higher on the 1948 key than on the 1938 key. For instance, 217

[2] See footnote 1.

Table 4

Letter Grades of 1048 Psychologists on the 1938 and 1948 Strong
Vocational Interest Blank Psychologist Keys

		1948 Psychologist Key							
		C	C+	B—	B	B+	A	Total	Per Cent
	A			3	3	21	497	524	50*
	B+			2	8	34	114	158	16*
1938 Psychologist Key	B		2	3	22	45	76	148	14
	B—	1	1	10	24	32	19	87	8
	C+	2	2	13	12	22	3	54	5
	C	8	10	22	17	7	13	77	7
Total		11	15	53	86	161	722	1048	
Per Cent		1	2	5	8	15*	69*		100%

* The difference between the 66% who score A or B+ on the 1938 key and the 84% who score A or B+ on the 1948 key is statistically significant (p. < .01).

psychologists score B or lower on the 1938 psychologist key and B+ or A on the 1948 key while only 16 score B or lower on the 1948 key and B+ or A on the 1938 key. These results seem to indicate that the two keys cannot be considered equally valid, and since the 1948 key is based on a larger and more up-to-date sample it is recommended that the 1948 key be substituted for the 1938 key.

New Sub-Keys

Sub-keys were constructed for four of the largest field sub-groups: experimental, clinical, guidance, and industrial psychologists. Since these keys were intended for the guidance of advanced psychology students who might be undecided as to the field in which they should specialize, it seemed that keys which contrasted the interests of each sub-group with the interests of other psychologists would be most useful. Consequently instead of Strong's professional men-in-general group, the 1948 sample of psychologists-in-general was used as a reference point. Actually the reference point was a slightly shifting one as the criterion group in each instance was subtracted from the total group. Strong's method of weighting responses was again followed.

The four criterion groups vary in size from 108 to 256. Although these groups are not as large as Strong recommends, they are large enough to assure reasonably accurate results. Strong has found that keys based on 250 cases are likely to differ from keys based on 500 cases by

only one or two standard scores while keys based on 100 cases are likely to differ from those based on 500 cases by between two and eight standard scores.[3]

The extent to which the sub-keys differentiate the four criterion groups from psychologists-in-general is presented in Table 5. Before analyzing these results, however, it may be well to consider some of the factors which have affected the degree of differentiation secured.

In the first place, since all available members of each criterion group were needed in constructing the key for that group, it was not possible to rescore an independent sample. The separation we have obtained is consequently greater than would be expected for other samples. Secondly, the shifting reference point which we chose to use in this study also tended to increase differentiation between criterion and reference groups.

Table 5

Power of Four Strong Vocational Interest Blank Sub-keys for Psychologists to Differentiate Criterion and Reference Groups

Sub-key	Criterion Group			Reference Group		Difference Between Means in Criterion S.D. Units	Percentage of Criterion Group Exceeding Reference Group Mean
	N	Mean Raw Score	Standard Deviation	N	Mean Raw Score		
Clinical	221	47.5	20.9	827	23.0	1.2	85.1
Experimental	256	42.9	38.8	792	-10.6	1.4	88.7
Guidance	115	50.0	37.6	933	- .1	1.3	91.3
Industrial	108	35.7	29.9	940	- 3.9	1.3	88.0

On the other hand, the fact that all psychologists were forced into one of the nine field classifications has tended to give us less separation than if we had used only those psychologists who clearly fell into one of the classifications. Finally, it should be pointed out that, other things being equal, the smaller the criterion group, the greater will be its separation from other psychologists. This is true because sampling errors can be capitalized upon to a greater degree with a small group than with a large group.

The results reported in Table 5 indicate that these four sub-keys have about equal differentiating power. In all four instances the reference and criterion group means are about 13 standard scores, or 1.3 standard deviations apart, and over 85% of the members of the criterion

[3] Strong, E. K. *Vocational interests of men and women.* Stanford University Press, 1943, pp. 645-646.

groups exceed the means of the reference point groups. These four keys do not secure the extremely sharp separation between criterion and reference point groups obtained by the 1948 psychologist key, but it would seem that they have sufficient differentiating power to warrant their use. Moreover, the four sub-keys have the advantage of being based entirely on recent data while the 1948 psychologist key, since it contrasts the interests of present day psychologists with the interests of a 1938 professional men-in-general group, may be somewhat out of date.

An analysis of the response weights assigned to items for the four sub-keys, presented in Table 3,[4] indicates that clinical psychologists are differentiated from other psychologists by greater artistic, literary, teaching, verbal, and social service interests. Experimentalists have stronger interests in physical science, mathematics, and mechanical work. Guidance psychologists have a stronger preference than others for interviewing, service to others, personnel work, and writing. Industrial psychologists are distinguished by their business interests.

Correlation between Keys

The relationships among the five new keys developed in this study are shown by the following correlations based on a sample of 216 psychologists not included in any of the four sub-key criterion groups. (To be exact these correlations are not between keys but between the scores of individuals on pairs of keys.):

	Psychologist Key	Clinical Key	Experimental Key	Guidance Key
Clinical Key	$.30 \pm .09$			
Experimental Key	$.25 \pm .09$	$-.52 \pm .07$		
Guidance Key	$-.32 \pm .08$	$.28 \pm .09$	$-.82 \pm .03$	
Industrial Key	$-.36 \pm .08$	$-.13 \pm .09$	$-.37 \pm .08$	$.54 \pm .07$

Care must be used in interpreting these correlations. The positive correlation between the clinical key and the psychologist key means that clinical psychologists tend to differ from other psychologists in somewhat the same way that psychologists-in-general differ from professional men-in-general. The positive correlation between the guidance and industrial keys means that guidance psychologists tend to differ in the same way from non-guidance psychologists as industrial psychologists differ from non-industrial psychologists.

[4] See footnote 1.

Conclusions

It is suggested that better guidance for potential psychology students can be given if the 1948 psychologist key is substituted for the present (1938) key, and if consideration is also given to the profile scores for psychologists-in-general now available.[5] It is also suggested that advanced psychology students who are undecided as to the field of psychology in which they should specialize, should take advantage of the four new subkeys and the profile data for **13** different kinds of psychologists developed in this study.

[5] The five keys developed in this study have been approved by Professor E. K. Strong, Jr. Interest blanks sent to Stanford University are now scored on the 1948 rather than on the 1938 psychologist key. Engineers Northwest (100 Metropolitan Life Building, Minneapolis 1, Minnesota) is equipped to machine score both the 1948 psychologist key and the four psychologist sub-keys on request.

Kuder Interest Patterns of University Business School Seniors

Robert H. Shaffer

Indiana University

This paper reports the findings of an analysis of the mean raw and percentile scores made on the Kuder Preference Record by Indiana University School of Business seniors in the classes of 1947 and 1948. Particular reference is made to the interest patterns which were found to be characteristic of students majoring in the various curricula.

Procedure

The Preference Record was administered to the 975 men and 205 women students in the graduating classes of 1947 and 1948. As a vocational interest inventory it is widely used at the present time for vocational counseling and, to some extent, in selection and placement. It yields scores in nine general areas of preference: mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, and clerical.

The mean raw scores with standard deviations were calculated for each of the total groups by sex and for various sub-groups classed by college major. The means of the students grouped by major were compared with the means of the corresponding scales of the total group by use of the critical ratio technique.

Findings

Table 1 gives the mean raw scores in the nine areas of the Kuder Preference Record for the total group of business school seniors and for subgroups divided by major subject. Table 2 gives the percentile ranks based upon the mean raw score as given in the published norms for the test. Table 3 gives the raw scores for the women students and Table 4, the equivalent percentile scores.

The comparison of the mean raw scores revealed that the business school seniors studied had varied, and in many cases, markedly different interest patterns (Table 1). The accounting and advertising majors had scores significantly different from the total business school group in all nine areas. All of the advertising scores were significant at the 1%

Table 1

Kuder Preference Record Mean Raw Scores by Major for Senior Men in Indiana University School of Business

Major	N		Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Ser.	Clerical
Total Group	975	M	64.8	40.5	54.7	97.0	39.7	51.4	19.0	69.4	61.1
		SD	18.2	14.8	13.5	19.8	12.7	14.6	9.3	17.0	14.3
Gen. Business	179	M	62.3	40.9	54.1	96.2	40.8	51.8	19.8	69.5	60.8
		SD	18.1	12.7	13.6	19.4	12.5	14.9	8.6	16.4	13.6
Accounting	217	M	67.9*	57.1**	58.3**	84.3—	37.6—	48.4—	17.1—	64.7—	70.3**
		SD	17.4	8.9	11.6	15.8	12.1	13.9	9.5	15.2	13.5
Finance and Banking	37	M	65.7	43.9	53.5	92.9	37.8	56.3**	20.0	67.6	61.1
		SD	17.9	13.1	12.0	24.5	13.3	12.5	10.1	18.5	14.4
Management	138	M	68.6*	37.4—	53.9	96.2	37.3—	49.8	17.6	76.2**	61.5
		SD	19.5	12.4	13.2	16.4	12.0	13.4	8.9	18.2	13.7
Advertising	107	M	59.7—	30.9—	49.9—	104.0**	45.8**	57.4**	21.5**	64.4—	57.0—
		SD	17.2	11.8	14.9	18.4	13.9	16.5	9.3	15.3	13.3
Retailing	109	M	65.4	33.1—	54.7	105.1**	40.7	51.8	17.8	70.0	56.3—
		SD	17.5	11.3	12.1	17.0	12.1	13.6	9.1	17.5	12.5
Sales	136	M	64.0	30.7—	52.3	112.3**	38.8	50.8	20.4	72.5*	55.6—
		SD	18.2	9.8	13.4	12.2	11.3	14.1	8.9	15.2	11.6

Note: In Tables 1, 2, 3, and 4, * indicates a positive difference significant at the 5% level, ** indicates a positive difference significant at the 1% level, – indicates a negative difference significant at the 5% level, and — indicates a negative difference at the 1% level.

Table 2

Mean Kuder Preference Record Percentile Scores of Senior Men by Business School Major

Major	N	Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Ser.	Clerical
Total Group	975	25	72	21	96	29	65	57	70	73
Gen. Business	179	22	74	20	96	31	66	60	70	72
Accounting	217	29*	97**	27**	88—	24—	56—	51—	57—	89**
Finance	37	26	81	18	95	24	76**	61	66	73
Management	138	31*	61—	20	96	23-	60	53	82**	74
Advertising	107	20—	36—	13—	98**	47**	79**	66**	56—	63—
Retailing	109	26	45—	20	98**	32	66	52	67	66—
Sales	136	24	35—	17	99**	27	62	62	76*	57—

Table 3

Kuder Preference Record Mean Raw Scores by Major for Senior Women in Indiana University School of Business

Major	N	Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Ser.	Cleri-cal
Total Group	205 M	47.5	33.2	47.1	80.0	51.2	54.3	22.3	76.9	67.3
	SD	14.8	13.7	15.4	18.1	16.5	14.7	8.6	18.0	18.1
Gen. Business	32 M	46.4	40.2**	50.5	76.5	49.6	54.8	22.4	72.5	72.2
	SD	14.9	13.1	14.0	17.8	13.6	15.3	6.9	18.6	13.2
Management	18 M	46.2	32.6	43.9	84.6	46.4	51.5	21.7	84.6	68.4
	SD	16.9	12.5	12.9	17.4	15.1	11.2	6.8	16.5	12.7
Advertising	30 M	49.2	24.3—	42.0	87.4*	65.1**	57.1	23.2	66.0—	59.3-
	SD	15.3	8.0	13.6	17.8	15.5	12.9	7.9	19.6	10.1
Retailing	24 M	48.7	31.7	43.3	89.7**	53.2	52.0	22.0	73.5	63.0
	SD	16.3	10.2	12.7	15.1	15.4	14.4	7.5	15.8	12.9
Secretarial	26 M	41.5-	30.6	44.9	77.5	48.8	53.4	24.5	79.7	81.9**
	SD	13.6	10.3	17.0	14.9	15.3	14.0	8.0	14.7	16.4
Com'l Teacher	30 M	44.6	35.8	47.0	71.9-	44.6—	51.5	20.6	87.4**	75.5*
	SD	10.5	11.9	14.1	16.1	11.8	11.1	8.9	14.5	17.0

Table 4

Mean Kuder Preference Record Percentiles of Senior Women by Business School Major

Major	N	Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Ser.	Cleri-cal
Total Group	205	48	68	40	87	52	66	46	45	59
Gen. Business	32	45	88**	50	80	48	66	46	35	70
Management	18	45	65	31	92	40	59	42	62	62
Advertising	30	52	35—	21	95*	79**	73	50	23—	40-
Retailing	24	51	62	30	96**	59	61	45	37	50
Secretarial	26	31-	59	34	82	45	64	55	52	81**
Com'l Teacher	30	39	76	40	66—	36—	58	38	68**	76*

level with positive scores in the persuasive, artistic, literary and musical areas and with negative scores in the mechanical, computational, scientific, social service and clerical areas. The accounting group had relatively high scores in the computational, scientific, clerical and mechanical areas with relatively low scores in the persuasive, artistic, literary, musical and social service fields.

Other patterns found included the following: a relatively high literary interest by the finance group; positively significant social service and mechanical scores for the management with negatively significant scores in the computational and artistic areas; one positively significant score, in the persuasive area, for the retailing group with negatively significant scores in the computational and clerical areas; and a persuasive-social service pattern for the sales group with negative scores in the computational and clerical areas.

It is important to note that these observed patterns are based on differences in the mean raw scores and not on percentile scores. For example, the mean raw score of the advertising group in the artistic area was significantly higher at the 1% level than the mean score of the total group, yet the corresponding percentile score was only 47 (Table 2). Similarly, the mean score of the accounting group in the persuasive area fell at the 88th percentile, yet it was significantly lower than the mean score of the base group.

Similar but less marked patterns were found for the women seniors (Tables 3 and 4). Judging from the number of significant scores, the women students majoring in advertising had the most definite interest pattern with significantly high artistic and persuasive scores and low computational, social service and clerical scores. Students studying to be commercial teachers had four significant scores, high social service and clerical and low artistic and persuasive. Contrasted with this pattern was that of the secretarial students with a high score in the clerical area and a low score, significant at the 5% level, in the mechanical area. As in the case with the men, the women retailing majors had only one high score, in the persuasive area.

The general business majors had the highest score in the computational area and the lowest score in the persuasive area of any of the subgroups. Their score in the clerical area was third high, being below the secretarial and the commercial teacher groups. Thus their pattern followed very closely the pattern of the male accounting majors.

An extremely high persuasive percentile score, based upon the general norms, was found to characterize all of the groups with the possible exception of the commercial teacher group (Tables 2 and 4).

Summary

The Kuder Preference Record revealed significant differences in the interest patterns characterizing senior students majoring in the various curricula in the Indiana University School of Business. In practically every case the interest patterns for the various groups followed those set

up for related occupations by the test manual. The need for establishing local norms and for analyzing percentile and raw scores carefully was emphasized by the deviation of the business school group from the general group used for establishing the published norms.

The findings indicate that the Kuder is a useful tool in assisting students to choose a major within a school of business.

An Analysis of Certain Factors in Serious Accidents in a Large Steel Plant *

John B. Whitlock, Jr. and Clarke W. Crannell

Miami University, Ohio

The group selected for study was comprised of the 100 most recent accident reports, beginning with the date upon which the study began. Starting with a case dated March 14, 1947, each accident was taken in reverse chronological order until 100 had been recorded. Case number 100 is dated February 20, 1944. The 100 accident records obtained in this manner thus include all major accidents at the Armco Steel Corporation's Middletown Division over a period of approximately three years.

For comparison with these cases, a "control," or accident-free group of two hundred cases was selected from men and women who had worked on the same jobs during the same period of time, but who had not had a major accident in that time. From the accident group reports a list of jobs on which accidents had occurred and the frequency of accidents on each was compiled. The foreman of each of the departments which was listed as having had one or more accidents in the three-year period was then contacted, and from these men the names of two accident-free men for each accident case listed was obtained. For example, the Repairmen Section of the Maintenance Department was found to have the greatest frequency, with twelve accidents; so the names of 24 accident-free men were obtained. In the case of the Masonry Department, where one bricklayer had been injured, the names of two bricklayers who had not been injured were obtained. This procedure yielded a group of 200 men and women who had worked without accident on the same jobs during the same period of time as had the 100 men and women listed as accident cases.[1]

For all individuals, the following information was transcribed from the company records to a specially prepared data sheet: (1) name; (2) check number; (3) age; (4) marital status; (5) dependents; (6) average weekly wage; (7) World War II veteran; (8) physical rating—as decided

---

[1] Because one job on which there had been an accident has been since discontinued, the control group actually consists of 198 cases.

by the company doctor at the time of employment; (9) height; (10) weight; (11) blood pressure; (12) vision; (13) test scores,—a. Otis Test of Mental Abilities; b. Bennet-Frye Mechanical Comprehension; c. Bernreuter Personality Inventory; all six percentile scores;—(14) education; (15) company service; and (16) job service.

For the accident cases, the following additional data were transcribed on the record sheet: (17) date of accident; (18) day of week; (19) time of day; (20) amount of shift worked; (21) whether doing usual job; (22) job when accident occurred; (23) classification of accident,—a. days lost, b. part of body injured, c. type (crush, burn, cut fracture, etc.), d. disability (temporary total or partial, permanent total or partial, or death); (24) description of accident; (25) how it happened; (26) why it happened; (27) cause; (28) responsibility; and (29) previous major accidents, if any, and classification.

Because the records of height, weight, blood pressure, and vision had often been made months or years prior to the accident, it was not considered feasible to include these variables in the investigation and they were not considered.

The test scores (Item 13 above) were not available on all cases because testing did not begin until 1942 and many of the men with whom this study is concerned were in the employ of the corporation prior to that date. Among the accident cases, 47 individuals had taken two or more of the tests. Among the control individuals, 62 to 65 had taken two or more of the tests.

The company reports of major accidents are divided into three groups on the basis of responsibility. This responsibility is determined by an accident investigation committee which meets as soon as possible after the occurrence of the accident. In the present study, the total group of accident cases was divided on the basis of the committee's decision into the following subgroups: A—those who were totally responsible for their accidents; B—those who were jointly responsible for their accidents; C—those who were injured by the action or lack of action of someone else.

Results

Non-test Data. No item of non-test data was found to differentiate significantly between the accident and accident-free groups. This was found true not only for each subgroup among the accident cases as divided according to responsibility, but also for certain subdivisions of the control group made on the basis of length of company service. Table 1 summarizes these non-test data.

Mechanical Comprehension and Otis Test Scores. These two tests failed to reveal any significant relationship to the occurrence of accidents.

Inspection of the Otis Test scores seemed to indicate that these scores increased in the control group with age and company service. An analysis of variance applied to these scores in terms of Age and Service demonstrated that the variance was significantly related in approximately equal amounts to both of these variables. The possibility is therefore not excluded that, were it possible "experimentally" to equate the accident and control data with regard to age and service, a relationship to Otis Scores might be found.

Table 1
Mean Values for Non-test Data

Group	Age	Dependents	Weekly Wage	Company Service (days)	Education
Accident A	39.0	2.00	$52.92	3255.5	7.84
Accident B	31.8	1.87	48.91	2168.9	8.94
Accident C	34.9	2.00	49.42	3251.4	8.73
Control	41.8	2.82	57.80	5128.0	7.88

Bernreuter Personality Inventory. The company records available to the present writers presented the Bernreuter scores in terms of percentiles. For the purpose of statistical computations, these scores were converted into T-score equivalents. Table 3 shows the means and standard deviations for each of the six Bernreuter scales. The writers are fully aware of the hazards in interpreting the scores of personality measures which have been obtained in industrial situations. It is quite probable that applicants for a job will give what appears to them to be a "correct" answer on such an inventory, rather than an answer which may reveal their real opinions of themselves. It should therefore be kept in mind when reading the following discussion, that the terms alluding to personality traits are employed for convenience in identifying the scales according to the system employed by the author of the inventory, and it

Table 2
Mean Test Scores by Accident Groups

Accident Group	Mech. Compr. Score		Otis Score	
	N	M	N	M
Accident A	9	30.9	10	82.9
Accident B	17	33.3	17	93.4
Accident C	21	31.3	21	87.8
Control	62	34.0	64	84.5

should not be implied that the writers believe the scores to be truly representative of any unitary personality trait.

Some of the mean differences between accident and control groups, computed from data summarized in Table 3, were found to be significant when Fisher's t test was applied. On the Bl-N scale (neurotic tendency) the mean T-score for Accident Group A was significantly lower (less "neurotic") than the control group at the 5% level of confidence (t: 2.06). The same comparison between control group and Accident Groups A and B combined yielded a significant difference at the 1% level. Closely comparable results were obtained for the B3-I scale (introversion-extroversion). One further significant difference was found: between Accident Groups A and B combined and the control group for the Fl-C scale (confidence in oneself), the mean for the combined accident groups being significantly lower (more "self-confident") at the 1% level.

Table 3
Mean T-Score Equivalents of Bernreuter Personality Inventory

Group	N	M	SD	M	SD	M	SD
		B1-N		B2-S		B3-I	
Accident A	10	40.2	6.78	49.8	6.34	39.0	6.00
Accident B	17	36.9	8.02	48.9	8.02	36.5	8.84
Accident C	20	43.1	8.96	50.6	5.95	41.4	7.86
Control	65	45.3	7.61	48.0	6.60	43.4	6.60
		B4-D		F1-C		F2-S	
Accident A	10	52.5	5.41	42.9	7.23	42.6	6.40
Accident B	17	52.9	7.12	41.5	9.39	38.7	6.16
Accident C	20	51.6	6.43	43.1	9.16	48.9	6.07
Control	65	49.1	6.69	46.4	8.35	44.1	6.22

An inspection of the means in Table 3 reveals that among the first four scales the accident groups present alternately higher and lower values as compared with the control group. The writers examined the individual data for each scale, and discovered that individuals who were considered totally or partially responsible for their accidents seemed to have wider discrepancies in their scores from one scale to the next than did those individuals who were accident-free or not considered responsible for their accidents. This feature of the data was most evident when the B3-I and B4-D (dominance-submission) scales were compared. By subtracting the B3-I score from the B4-D score for each individual, difference scores were obtained which were preponderantly positive for Groups A and B—only one in Group A and three in Group B were zero or negative. On the other hand, 22, or one-third, of the cases in the control

group showed zero or negative differences between these measures. The mean difference scores for Groups A, B, C and control were 14.0, 16.5, 9.8 and 5.9, respectively. The means for Groups A and B are each significantly different from the control group at the 5% level or better. While there is considerable overlap among difference scores of the various groups, there does appear to be a significant trend for accident cases to have higher B4-D scores ("dominant") and lower B3-I scores ("extroverted").

Summary

From the accident report records of the ARMCO Steel Corporation, Middletown, Ohio, data were collected on 100 accidents which occurred over a three-year period ending in March 1947. Besides the accident data, personal data and test scores (where available) were obtained for these individuals and for a control group composed of two accident-free employees for every accident case. A statistical analysis was made with the following results:

1. In this particular study, none of the "non-test" data was found to differentiate between the accident and control groups.

2. Mechanical Comprehension and Otis Intelligence Test scores were not found useful for accident prediction purposes, but the Otis was found to be greatly affected by age and company service.

3. Three of the Bernreuter Personality Inventory Scales seemed to differentiate to some degree between accident and accident-free groups. In the nomenclature of this Inventory, the accident cases appeared less "neurotic," less "introverted" and more "self-confident." Especially noteworthy was a tendency among the accident cases to have high B4-D (dominance) scores and at the same time low B3-I (extroversion) scores. In view of the wide overlap among groups in these measures, and also considering the rather speculative nature of personality assessment by means of inventories, evidence of this sort can hardly be taken as a reasonable basis for exclusion from employment. Nevertheless it does suggest that employees who conform to the "personality pattern" of such accident cases should be subject to closer scrutiny when placed on jobs involving the possibility of serious injury.

The small number of cases studied here, and the limitation of the study to one industry, necessarily limit our conclusions to suggestions for further study. It is definitely believed that the variability among Bernreuter Personality Inventory scores would make a profitable subject for further study with a large number of cases.

Visual Performance and Accident Frequency

Joseph Tiffin

Purdue University

and

B. T. Parker and R. W. Habersat

Bausch & Lomb Optical Company

The importance of adequate vision as a safety factor has received increased recognition from industrial safety engineers in recent years. Numerous studies have pointed out the desirability of requiring industrial job applicants to meet minimum safety visual requirements at the time of employment. In some instances these studies have shown decreases in injury frequency following the introduction of such employment standards.

This paper summarizes the results of a recent experiment in a light manufacturing industry. The data confirm previous evidence that low visual performance and injuries frequently go hand in hand. In this case the use of a minimum visual safety standard for employment on all factory jobs probably materially reduced injury frequency and compensation costs.

Procedure

The medical and accident records in a large optical company were examined and employees who had experienced three or more injuries [1] in the previous 18 month period were identified. These were considered "high frequency of injury" employees compared with the average experience in this plant. It was planned to compare the visual performance of this group with that of an injury free group to see whether there were any significant differences.

Age, education, experience and the job hazards have all been shown to be related to injury susceptibility (1, 3, 5, 6 and 8). In setting up the injury free or control group these factors were therefore carefully controlled. This was done by matching each employee in the accident group with an employee on the same job, and having the same age,

---

[1] Types of injuries included were: fractures, bruises or contusions, sprains, cuts, abrasions, slivers, etc., or any other types of injuries that could possibly have been caused by low visual performance. Such injuries as hernias and back strains were not considered for the purposes of this study.

education and experience, but who was "accident free" during the 18 month period used for the study. This group of "accident free" employees made up the control group.

It was also necessary to make sure that none of those included in the two groups had received professional eye attention during the 18 month period of the study since such eye care undoubtedly would improve visual performance which might in turn have reduced the employee's injury susceptibility. Since complete eye examination records were available on most employees in the plant's own eye clinic, it was possible to determine in most cases in advance of testing which employees had received professional eye care in this time. As an additional check after the pairing was completed and the testing begun, each subject was questioned carefully concerning the last date eye care had been obtained. Where either one of the pair had received professional eye attention during the previous 18 month period, the pair was dropped from further study. After this last control had been applied 42 matched pairs remained, a total of 84 employees participating in the experiment.

The visual performance of each individual was tested on the Bausch & Lomb Ortho-Rater (4), a precision instrument measuring the visual skills listed in the first column of Table 1.

On completion of the visual performance testing, means were computed for the raw scores of each group on the vision tests. Critical ratios were also computed to determine the statistical significance of any existing differences (2). Both means and critical ratios are shown in Table 1 in comparison with those obtained in a similar study conducted in a heavy industry (7).

Results

Table 1 shows that on the average, the injury free group of employees had superior visual performance.

In three visual skills—acuity worse eye (near vision), acuity, right eye (near vision) and color perception, differences exceeded the 5% level of significance. In Stump's study (7) the critical ratios were high for several distance visual skills, but very low for near visual skills. This may be due to the fact that on the jobs in a heavy industry, distance visual skills are more frequently required in performing the job. On the other hand, for the jobs covered by the present study in the optical industry, near visual skills are essential and are therefore important for safety.

These studies would seem to indicate that somewhat different patterns of visual skills may be required for safety on various jobs in different industries. Each plant should probably conduct its own investigation

Table 1

Mean Ortho-Rater Visual Performance Scores and Critical Ratios of Injury Free and High-Frequency-of-Injury Employee Groups in Two Independent Studies

Tests	Study No. 1 (Heavy Industry, see Stump (7))			Study No. 2 (Light Industry, present study)		
	A Injury Free Group Mean Score	B High Frequency Group Mean Score	Critical Ratio M_A-M_B	C Injury Free Group Mean Score N = 42	D High Frequency Group Mean Score N = 42	Critical Ratio M_C-M_D
1. Vertical Phoria (Far)	5.56	5.25	1.02	5.49	5.48	0.00
2. Lateral Phoria (Far)	7.44	6.72	1.25	8.29	8.00	0.51
3. Acuity, Both Eyes (Far)	10.81	10.14	1.69	9.90	9.68	0.68
4. Acuity, Right Eye (Far)	9.89	8.83	2.26**	9.67	8.43	1.95
5. Acuity, Left Eye (Far)	9.86	8.67	2.08**	9.69	9.26	0.86
6. Acuity, Better Eye (Far)	10.53	9.72	2.21**	10.36	9.81	1.67
7. Acuity, Worse Eye (Far)	9.25	7.78	2.56**	9.00	7.88	1.73
8. Depth (Far)	5.64	3.53	3.27*	5.55	4.74	1.41
9. Color (Far)	4.44	4.25	0.60	4.93	4.43	2.14**
10. Acuity, Both Eyes (Near)	11.08	10.69	0.87	8.93	8.38	1.38
11. Acuity, Right Eye (Near)	9.03	8.58	0.70	8.86	6.88	2.95*
12. Acuity, Left Eye (Near)	9.61	9.28	0.57	9.00	7.93	1.59
13. Acuity, Better Eye (Near)	10.00	10.03	0.05²	9.62	8.95	1.52
14. Acuity, Worse Eye (Near)	8.64	7.86	1.24	8.24	5.88	3.36*
15. Vertical Phoria (Near)	4.61	4.61	0.00	4.60	4.50	0.32
16. Lateral Phoria (Near)	6.33	7.14	1.02³	7.83	7.83	0.00
Mean Age	—	—	—	41.8	41.8	—
Mean Education (Grades Completed)	8.53	9.00	0.90⁴	8.21	8.24	.08⁵
Mean Experience (Years on Job)	—	—	—	9.45	7.67	1.08

* Significant at 1% level or less.
** Significant at 5% level or less.
²,³,⁴ These critical ratios were computed from M_B-M_A.
⁵ This critical ratio was computed from M_D-M_C.

on a fact finding basis to determine what visual skills are required for safe operation, and establish visual safety standards accordingly.

Further examination of Table 1 reveals that in general the level of visual performance of the workers studied in the first experiment was considerably higher than that of the workers in the present study. This could be due to a difference in the average age levels of the individuals sampled in each study, to job differences, or differences in the sources of employees (8).

Value to Management

Since 1943 visual safety standards have been in use in this company for selection and placement of new applicants, and referral of present employees for eye care. During this entire period compensable accident costs have steadily decreased compared with the previous four year period. Due to the fact that changes in safety procedures are constantly taking place in all progressive plants interested in safety, it is, of course, difficult to ascertain how much, if any, of a decrease in injuries or accident costs can be attributed to one specific part of the overall program. Since the addition of the vision program was the only major change in the safety procedure in the plant studied during the period of this investigation, it is probable that it played a substantial role in the reduction of direct compensation costs by an average of $16,600 per year, which occurred over a four year period. Studies of small groups of employees in the factory have shown a definite decrease in injuries after professional eye attention. Further investigations of the effect of eye care on reduction of injuries are now in progress.

References

1. Chambers, E. G. A preliminary inquiry into the part played by character and temperament in accident causation. *J. ment. Sci.*, 1939, **85**, 115–118.
2. Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill Book Co., 1936, 548–549.
3. Henig, M. S. Intelligence and safety. *J. educ. Res.*, 1927, **16**, 81–87.
4. Jobe, F. W. Instrumentation for the Bausch & Lomb industrial service. *Bausch & Lomb Magazine*, 1944, **20**, 6–7, 14–15.
5. Lipmann, O. *Unfallursachen und unfallbekampfung.* Berlin: 1925.
6. Schmitt, E. Unfallaffinität und Psychotechnik im Eisenbahndienst. *Industrielle Psychotechnik*, 1926, **3**, 144–153, 364–365.
7. Stump, N. F. A statistical study of visual functions and safety. *J. appl. Psychol.*, 1945, **29**, 467–470.
8. Tiffin, J. *Industrial psychology.* New York: Prentice-Hall, Inc., 1947, **13**, 425–430, 435–439, 443–444; and 228–230.

Attention and Involuntary Movement

Austin S. Edwards

University of Georgia

In a situation such as automobile driving, how important is involuntary movement? To what extent is uncontrolled movement different with fixed and with shifting attention? To what extent is two-armed driving more steady and controlled than is one-armed driving? It is not sufficient to know that certain conditions are important for safety. It is desirable to know as accurately as possible to what extent certain conditions modify the control of activities in the total behavior of individuals who are involved in skilled work or in what may be dangerous occupations. Although inferences cannot be made directly from laboratory experiments to such activities as automobile driving, it is believed that the following experiments may throw some light upon the problems.

Two experiments have been performed for the purpose of discovering quantitatively the effect of certain conditions upon involuntary movement. The involuntary movement chosen was finger tremor, arm extended with no rest, since that can be accurately measured in three dimensions with the writer's finger tromometer.[1] The two conditions especially used were distractions with attention held upon a fixation spot without any shifting, and, second, the effect of distractions when the attention shifted.

Experiment 1   Involuntary Movement with Fixed Attention

This experiment was performed in the dark room with the light dimmed to one foot candle. The distractions were an automobile horn from a Ford automobile, actuated by six volts, and the bright light from a Plymouth automobile, actuated by a current of six volts. The horn was in a box about five feet in front of the S, where it could not be seen. The light was placed about ten feet in front of the S slightly to the left and arranged so that it could easily be moved in any direction and shone into the eyes of the S. The finger tromometer was on the table directly in front of the S and a fixation spot about six feet in front so that the S could look over the top of the tromometer and fixate the spot.

[1] Edwards, A. S.   The finger tromometer.   *Amer. J. Psychol.*, 1946, 59, 273-276.

Procedure. Standard procedure was used with the tromometer, the *S* being allowed to rest before the measurements began, and the measurements consisting of the sum of the three readings—front-back, right-left, up-down. The time of measurement was thirty seconds.

The first control measurement was made before any distractions were used, and then measurements were taken while either the light or the horn, or both, were used as distractions. Following these three experimental measurements, a second control measurement was made. When the light was directed into the eyes of the *S*, or the horn was blown, the time consumed by the distraction was approximately twenty of the thirty seconds. Order of stimuli was varied so that each stimulus—light, sound, or both together—appeared with varying *S*s as either first, second, or third stimulus. The order of stimulation was recorded for each *S* so that not only could the results of each stimulating condition be studied, but also the effect of the stimuli as regards order, namely, first, second, or third, could be studied.

Subjects. One hundred *S*s, unselected college students, half men and half women, were used in this experiment. The ages were 17 to 25. The *S*s were asked whether they had been in any accidents, automobile or other, and notes were made as to such accidents, their number and seriousness.

Instruction. The following instruction was used: "Lean back comfortably; both feet on the floor; hold your hand as steady as you can, and watch fixation spot during the testing."

The sound of the horn was probably somewhat louder and more disturbing than is found in traffic, and the automobile light shining into the *S's* eyes was closer and probably brighter than is usually found in actual driving situations. *S*s sat about five minutes before the experiment began.

Since it might soon become known among students that sound and light distractions were being used, conditions were made as equal as possible for all *S*s by telling them at the beginning of each series of measurements that there would be a control measurement, light and sound distractions, one or both; after the last measurement with distraction another control measurement was taken to compare with the first control.

Results. Detailed results were worked out with means, medians, standard deviations, sigma of the mean, Q_1 and Q_3. Analyses were made for men and women separately, both for the light stimulus and sound stimulus, and both together, and for the first, second, and third stimulus. This permitted finding out first whether greater effect might be found in finger tremor because of the sound, the light, or both, and second, whether the first, the second, or the third stimulus had more effect.

It was somewhat surprising to find that on the average there was no statistically significant increase in finger tremor. Both stimuli together had no appreciable effect greater than one of the distractions alone. Whereas it might be expected that the first distraction might be more disturbing and the later ones less, or, that with the second and third distractions the *S* might become more upset, neither result appeared. With all the averages running from about 37 to not quite 43 mm., the greatest increase in finger tremor caused by the distractions was for the men 8 per cent and for the women 9 per cent. None of the critical ratios between the means was significant for the entire experiment, the highest being 0.66, which indicates results not much better than chance.

It might be expected that those students who had been in automobile wrecks or who had had serious traumatic experiences would be more disturbed than the others. No such evidence was found. Some of those who had been in the worst accidents had the lowest finger tremor throughout the experiment.

The only positive results that can be stated are in the cases of a few exceptional students who were very greatly disturbed and who showed greatly increased finger tremor during one or all of the experimental measurements. Of these *S*s, and considering those whose finger tremor was increased 50 per cent to more than 100 per cent, there were 16 men and 12 women.

Of the 16 men considerably affected by sound, light, or both distractions together, it appeared that both distractions together affected the men most and most frequently, 12 of the 50; sound affected very considerably 6, and light 4 of the 50 men;

Of the 12 women considerably affected by the distractions, only 5 were greatly affected by both distractions together; 9 by sound, and 6 by light. It is to be noted that some of the 16 men and some of the 12 women were greatly affected by more than one of the distracting stimuli.

Considering the order of distractions, first, second or third, for the 16 men and 12 women greatly affected, 6 men were most affected by the first distraction, 6 by the second, and 4 by the third. Of the 12 women, 5 were most affected by the first distraction, 6 by the second, and only 1 by the third.

Taking these cases and our averages for the 100 *S*s altogether, there is no evidence that two distracting stimuli cause more involuntary movement than one of them alone; or that a series of distractions cause more disturbance in involuntary movement than do a first or second disturbing distraction.

It may also be noted that there is no evidence of a general build-up of disturbing influence caused by a series of three disturbing stimulus

situations. This is, of course, not without exceptions in certain of the *Ss*. But the first control average was 38.29 mm. and the second control average following the distractions was only 39.46, an insignificant difference of only 1.18 mm.

Conclusions. It appears from this experiment with steadily fixed attention that, with certain exceptions (16 of the 50 men, and 12 of the 50 women), students selected at random and irrespective of experience with accidents showed on the average no statistically significant increase in finger movement under conditions which were assumed to be considerably distracting and might have been expected to be quite disturbing.

On the other hand, 32 per cent of the men and 24 per cent of the women had very considerable increases of involuntary movement.

Considering all of the cases and the specially disturbed *Ss* altogether, there is no evidence that the two disturbing stimuli were more disturbing than was one at a time; nor that the third distraction had any more effect than the first or second.

There was no build-up of disturbing effect since the first control experiments and the last showed no significant difference in average.

Some suggestion appears from this experiment to corroborate what we already know, namely, the importance of fixed attention in connection with motor control.

Experiment 2   Involuntary Movement with Shifting Attention

In this experiment conditions were changed in several ways. Preliminary tests were made with the same set-up described in Experiment 1, but with the instruction to *S* to shift attention and to look sideways during the measurements. The preliminary experiments indicated decidedly different results and led to the development of an experiment in which the steering wheel of an automobile was fastened to two units (front-back, right-left) of the tromometer. The steering wheel was mounted so that it was at the height and angle of the steering wheel in a car. The tromometer was placed so that the top of the steering wheel was between the two units of the tromometer that were used for measurements. Movement of the wheel thus moved one of the riders on the tromometer. Movement in the opposite direction moved the other rider. In this set-up the control experiment might have practically or almost zero recorded, because with both hands on the wheel it was possible to hold it very steady. Also, two riders instead of three were engaged. All measurements, both control and experimental, were thus very much reduced.

Procedure. *S* was given time to rest before the control measurements were taken. He was told that during the experiment he was to remain

as steady as possible, to keep his feet flat on the floor, and to keep his eyes on the fixation point unless told otherwise. He was to be as comfortable as possible in the chair which was placed at a comfortable position for the *S* for holding the wheel. The first control measurement was made with *S* placing both hands on the wheel and holding his attention steadily on the fixation spot, which was three feet in front of him. Time of measurement was thirty seconds. After *S* had rested, the first experimental measurement was taken; *S* was in the same position as in the control measurement, but after fifteen seconds he was told to look out of the window, which was six feet to his right. After four seconds *S* was told to look back at the fixation spot. The second control measurement was taken with only one hand on the steering wheel. *S* was told to use the hand with which he wrote, and to hold attention steadily on the fixation spot. The next experimental measurement was made with one hand on the wheel, but after fifteen seconds *S* was told to look out of the window. After four seconds *S* was told to look back at the fixation spot. The third experimental measurement was made by having *S* hold both hands on the wheel, but after fifteen seconds *S* was told to take the pencil which was handed to him by *E* and then to put his hand back on the wheel. The pencil was handed to him at a distance of two feet. There were thus two control and three experimental measurements.

Subjects. In this experiment there were 60 men, aged 18–35, and 40 women, aged 18–24, all college students selected at random.

Results. The results in this experiment are in direct contrast with those of Experiment 1. For the 100 students, with the men and women studied separately, the effect of shifting attention was great and consistent, all differences with definitely significant critical ratios, 2.526 to 9.71. For the men, the means showed an increase from the first control to the first experimental situation on the average from 3 mm. to 4.6 mm. The second control was 6.18; the second and third experimental ratings 9.05 and 14.48 respectively. For the women, the increase from control to first experiment was from 2.38 mm. to 2.85, and from the second control of 4.53 mm. to third and fourth experimental measurements 6.93 and 11.7. The smallest percentage increase was 40 and the largest percentage increase was 158. The average percentage increase for all *S*s was 80.9; for men, 82.25, for women, 79.5. See Figure 1.

But compared with the first controls (both hands on the wheel and fixed attention) other conditions showed increases of uncontrolled movement of 300 to 400 per cent. On the basis of the averages, and taking the first control series for the men and for the women, shifting the attention to reach and take a pencil increased involuntary movement for the men, 4.46 times, and for the women, 4.91 times. If so great increase of

involuntary movement takes place in the relative quiet of the laboratory, how much is to be found under the more disturbing conditions of actual automobile driving?

Reference to Figure 1 also shows a very significant difference in control, when only one hand is used instead of two. The uncontrolled movement is about twice as great.

Conclusions. With shifting attention and slight distraction (no bright lights or loud sounds), the increase in involuntary hand and arm movement was large, consistent, and statistically significant. Although one cannot draw conclusions directly from these experiments to what may be expected to happen in such a situation as driving an automobile, the question is raised as to how much danger exists in driving on account of
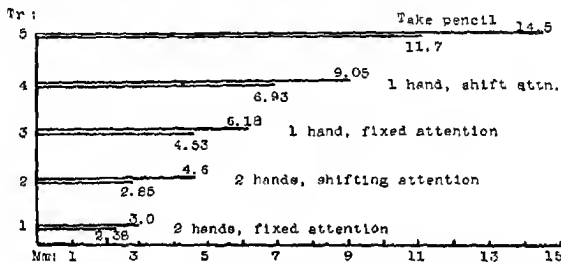
FIG. 1. Increased hand and arm movement with shifting attention. Tr indicates the series of trials: 1, the first control series; 2, the first experimental series; 3, the second control series; 4 and 5, the experimental series that followed 3. Horizontal lines are in mm., the upper line for men, the lower for women. There were 60 men and 40 women. The C.R.'s between means were all from 2.526 to 9.71.

involuntary movement over which the driver has no control. We may perhaps surmise that with the varying conditions of driving or of similar occupations, the varying amounts of monotony, emotional excitement, and various distractions might produce much more uncontrolled movement than we have found in the laboratory.

Considering the two experiments together it appears that hand and arm steadiness are relatively great with fixed attention; and that involuntary movement is relatively great when attention is shifting, especially when one is using one hand for a second operation.

It also appears that in a considerable number of cases involuntary movement of considerable amount may occur even when there is steadily fixed attention.

The great increase of uncontrolled movement when only one hand is used is of no small significance for automobile driving in traffic.

Practical Implications

Although it is clear to those who are informed that individuals suffering from certain abnormal conditions such as epilepsy and general paresis should not be permitted to engage in occupations where uncontrolled movements might happen, it has not been clear that there are so-called normal individuals and so-called normal situations that should require limitation of occupation in the interest of safety to self and to others. A more accurate and thorough knowledge of the involuntary movements that may occur with easily disturbed individuals and for practically all "normal" people under certain disturbing conditions may well find place in connection with our industries, automobile driving, and in other situations which demand steady neuro-muscular control. Too little is known and known accurately about the relation of involuntary movement to the total behavior pattern.

The question is forcibly raised: How dangerous is the one arm driving of automobiles?

Book Reviews

Burtt, Harold Ernest. *Applied psychology*. New York: Prentice-Hall, Inc., 1948. Pp. x+821. $7.35.

Burtt says: "This book is intended for two types of readers: the college student who has had an introductory course in psychology and desires orientation toward the science in its practical aspects, and the layman who is concerned with the general scope of applied psychology and possibly with a particular interest in some limited phase of it" (p. vii).

The author meets the needs of both audiences very well, for the materials are presented in a straightforward, factual way; and he comes to the assistance of those who have not had an introductory course in psychology by presenting in the second chapter a sort of "reader's digest" of such a course. Most statistical concepts are also avoided, and only those are employed which are essential for a first book in applied psychology.

After brief introductory remarks, followed by the resumé of general psychology, the remaining twenty-four chapters (consisting of 775 pages) seem to fall into the following categories: pseudo-psychology (astrology, spiritualism, graphology, phrenology, physiognomy, etc.); educational psychology, including individual differences and problems of learning; personal efficiency; vocational guidance; psychology in medical practice, and in psychotherapy; legal psychology, especially as related to problems of testimony and crime; business and industrial psychology, e.g., tests, industrial efficiency, fatigue, accidents; morale; the psychology of advertising, including consideration of advertising appeals and media; and a concluding chapter on "outlying fields."

In this last chapter are discussed: art, music, poetry, recreation and athletics, public speaking, writing, drama, moving pictures, radio, the press, public opinion, politics, religion, war, and peace. And all of this is done in 36 pages! This indicates the one general criticism of this textbook—that such important topics of applied psychology are so cursorily handled, while other subjects, such as crime detection and prevention, are dealt with at such length. Incidentally, the four chapters devoted to legal problems seem to be very largely a condensation of Burtt's book on *Legal Psychology* published in 1931.

Several important concepts are presented so briefly that the readers, of the kind to whom the book is addressed, will probably get erroneous notions of what the terms imply. Examples are retroactive inhibition (p. 127) and the psychoneuroses (pp. 258–260), although it must be ad-

mitted that it is difficult in any introductory or elementary book to do more than give readers a broad, general acquaintance with the field.

On the whole, the book is a good analysis of some of the more important areas of applied psychology. The addition of questions or other study aids at the end of each chapter would have been welcome for students; and it is unfortunate that no names of authors are given in the Index.

Stenart Henderson Britt

McCann-Erickson, Inc.
New York City

The OSS Assessment Staff. *Assessment of men: selection of personnel for the Office of Strategic Services.* New York: Rinehart and Co., 1948. Pp. xv+541. $6.50.

Whereas World War I produced one military personnel program in which psychologists played a crucial part, World War II produced four major and at least as many minor programs. Those of the Army Air Forces and of the Navy's Bureau of Personnel have been described in detail in volumes edited by Flanagan and his collaborators and by Stuit and his associates; the work of the psychologists in the Adjutant General's Office of the Army and in the Navy's Aviation Psychology Section of the Bureau of Medicine and Surgery has been described in scattered papers in the professional journals. And now Donald Fiske, Eugenia Hanfmann, Donald MacKinnon, James Miller, Henry Murray and others have made available in this volume a detailed account of the philosophy, methods, and results of the personnel selection program of the Office of Strategic Services.

This volume is as distinctive in literary style and as easy to read as any treatise by specialists in fantasy should be; it is as full of insights into personality structure and measurement as one would hope from personologists and clinicians; and it reveals the quick grasp of practical matters which was demonstrated time and again during World War II by academic and laboratory psychologists who turned to the practice of psychology. We might ask, What was the unique contribution of this group of World War II psychologists to personnel psychology?

The orientation of these psychologists was unique for, unlike the other groups working on personnel selection, known to this reviewer, this one included practically no personnel psychologists: they were laboratory men or clinicians, their interests were in personality theory and in psychotherapy. They therefore approached a personnel assignment with a freshness and originality, with a freedom from attachment to existing methods, which resulted in unusual creativity. Among the innovations

are a willingness to use subjective methods as a supplement to objective tests, the development of projective and situation tests, and the routine use of the intensive case conference. At the same time, lack of familiarity with the principles and practice of personnel psychology resulted in some unnecessary mistakes and in some "findings" which have long been familiar to workers in the field. To cite just one of several possible examples, despite an excellent analysis of the complexity and unreliability of their own criteria, they state (on page 397) that "If the job is running a lathe in a large factory the rate of piece-work production is a satisfactory criterion . . .," whereas various studies have shown that even as "objective" a criterion as output may be an unreliable and invalid index of a worker's success.

The organismic as opposed to the elementalistic approach to measurement is described and evaluated in detail, for perhaps the first time in a practical situation. Many of the techniques used by the OSS Assessment Staff were not new, and were not proclaimed as new, but they were thoroughly tried, revised in the light of experience, written up in detail, and statistically analyzed. The psychologist or personnel worker interested in the testing possibilities of social and practical situations, or in the use of the interview and of the case conference in evaluating leadership ability or ability to adjust to trying situations, will find many worthwhile suggestions in this account of the trial of these methods. This source is unusual among clinical studies, because the Assessment Staff was self-critical, checked for personal and systematic bias, and made all possible statistical analyses of their data as their work progressed.

The *organismic* method attempts to predict future behavior by inductive thinking from a set of observed facts to a conception (the hypothetical formulation of a personality), and then inductively to predict behavior in an anticipated situation, whereas the *elementalistic* approach attempts to predict future facts directly from observed facts.

It is especially interesting to compare the conclusions concerning the relative validity of the two methods reached by the authors wtih those of some workers in one of the other major personnel selection programs of the Armed Forces. The OSS psychologists state on page 227: "It is the contention of S (one OSS unit) that more can be learned of people by being with them as 'good Joes' than by testing them as professors." (Perhaps the reviewer will be pardoned for stating that this seems to him to be the discovery of someone who has previously done his testing as a professor, a laboratory man or clinician detached from the situation into which he injects his tests, rather than the wisdom of a person who has worked in a practical situation, who has used tests as only one method of gathering data, and who has had to be a "good Joe" all along in order

to do any kind of testing!) In contrast with this statement concerning the superiority of observational and projective techniques, one should consider the conclusions drawn from military testing experience by factor analysts such as Guilford in a recent article in the *Psychological Review*, proposing a selection procedure which is even more elementalistic and quantitative in its details than the traditional methods decried by the organicists of the OSS. There is no space in this review to explore these conflicting views, but they should be mentioned in juxtaposition, and some comparative data cited.

The OSS authors state that no adequate comparisons have been made between organismic and elementalistic approaches, but a few inadequate comparisons could be made. The AAF Aviation Psychology Program attempted to validate a number of clinical or organismic tests against success in flying training. Although time did not permit assessment by case conference procedures using all available information, single organismic techniques (e.g., interviews and Rorschach interpretations) were validated and proved to have no predictive value, while a number of single objective tests had validities in the .30s and .40s. In another study a psychiatrist and a psychologist interviewed cadets with borderline test scores and found that their clinical or organismic evaluations were no better than chance in predicting flying success in this group.

Despite their zeal for their procedures, the OSS staff frankly point out their defects ("Sometimes . . . we did not know who was deceiving whom . . ." p. 142) and, if anything, underrate their demonstrated validity. Concerning the validity of the assessment procedure, they state: "the final validity is a question mark" (p. 392). But this point bears consideration.

The validation procedure was made difficult by the scattering of OSS members over the world and by the lack of uniformity in assignments. It was made even more difficult by the fact that no attention had been paid to establishing criteria of success, a natural omission in the pressure of the early days of the war. Despite these and other problems, the average validity for the two principal assessment stations, for a sample of 171 men followed up and rated for job performance overseas by the Overseas Assessment Staff, was .45. There were no control groups to permit a comparison of the effectiveness of these with other methods, but taken by itself this is not a negligible validity coefficient.

A few interesting findings should be cited, to whet the appetites of possible readers. The motives of OSS volunteers were not generally ideological, but largely professional: trained men sought opportunities for worthwhile specialized experiences (p. 247; or do Americans tend to play down their idealistic motives?). Leadership is a relatively general

trait (p. 303). Cultural differences change the nature of situation tests: Chinese candidates ascribe leadership to their friends rather than to leaders (p. 352). Clinical psychologists are better able to diagnose than to predict, many apparent insights into personality being wasted because of lack of knowledge of the prediction situation (p. 430). There is a tendency to overevaluate outgoing persons with egocentric motives and low integrity, emotionally unstable individuals whose manifest behavior, is acceptable, and persons of low ability but possessed of goodwill and good social relations (p. 438 ff.). There were no systematic errors of underrating. Assessees gain considerable self-insight in the assessment process (p. 201). Traumatic experiences are common in the backgrounds of normal persons (p. 468).

Suggestions for future work are discussed in the last chapter. Again it may be worth pointing out the parallel with World War I, which brought ex-Army psychologists together in the postwar Scott Company and the Personnel Research Federation. After World War II some of the leading ex-Army psychologists founded the firm of Richardson, Bellows and Henry; a number of ex-Air Force psychologists launched the American Institute for Research; several established ex-Navy psychologists struck out on their own or went into business organizations; and now the final chapter of the OSS book reads almost like a request for a research grant. The Assessment Staff point out the possibilities inherent in their technique for the selection of executives and other key contact personnel (possibilities being capitalized by Selection Boards working with private industry and with government in Great Britain and in Australia), the unusual opportunities which such work affords for studying normal personality structure, and the ease with which assessment procedures such as these lend themselves to the training of junior psychologists in the observation, interpretation, and prediction of human behavior, that is, in the integrative processes with which most psychologists, both laboratory and clinical, have too little experience.

It is to be hoped that some far-sighted business or industrial concern, some foundation, or best of all, perhaps, some combination of the two, will be challenged by the prospect set forth by the Assessment Staff, and that organismic psychologists, both theoretical and applied, will have a real opportunity to develop and exploit the possibilities of their approach. Perhaps there will soon be a country house for psychological houseparties on the North Shore of Long Island or somewhere in Westchester! When the opportunity comes, it is to be hoped that the rules of procedure and recommendations for control and criteria made by the Assessment Staff will be carefully studied. The lessons learned and reported by the Assessment Staff should not have to be relearned by each group of

laboratory and clinical psychologists venturing into the personnel field. At the same time, their original and creative work should add considerably to the tools of the vocational psychologist and should broaden his understandings and deepen his insights. This is a valuable book, worthy of careful study by students of personality and by students of man at work.

Donald E. Super

Teachers College,
Columbia University

Escalona, Sybylle K. An application of the level of aspiration experiment to the study of personality. *Teach. Coll. Contr. Educ.* No. 937. Pp. viii + 132. Cloth $2.10.

A report of the study of the level of aspiration behavior of high school children is presented in this monograph. Although seventy-eight cases were originally studied, this report deals with the comparison of nineteen maladjusted and nineteen well-adjusted subjects. The level of aspiration techniques used were similar to Jucknat's, where the subject selects his task from a series of tasks of graded difficulty placed on a table, but also included a pre-arranged sequence of success and failure experiences such as Gardner used. In addition to the usual level of aspiration scores, decision time and voluntary discontinuation after failure were also measured. The method also included careful interviewing of the subjects following the test procedure.

Two of the interesting quantitative findings were concerned with decision time and voluntary discontinuation. It was found that those in the maladjusted group spent significantly longer time in deciding what task to attempt, and they showed a greater tendency to wish to discontinue the experiment following a failure when given an opportunity to do so. It was also found that those in the maladjusted group were more likely to lower their choices after failure than the subjects in the normal group. In general, the results are interpreted in the topological framework, and a number of stimulating hypotheses are presented.

One major limitation of the study is the small number of cases. A second limitation is concerned with the grouping of cases into maladjusted and well-adjusted. Some of the maladjusted cases were aggressive delinquents, others were withdrawn daydreamers. Part of the author's failure to find consistent results on some measures is probably a function of the heterogeneous nature of her maladjusted group. The author did fractionate her maladjusted group into sub-groups for some measures, but these groups were unfortunately too small to make meaningful statistical analyses. Other limitations of the method for the study of personality characteristics of individuals were the short number of trials, the method itself which discouraged, although it allowed, a repetition of the same

choice following either success or failure, and the loss of flexibility involved in the prearranged sequence of success and failures. All of these factors tended to limit the study of patterns of response and to provide only a brief sample of the subject's behavior in what might be considered a free-choice situation.

The author concludes that although no one-to-one correlation between personality characteristics and particular responses in the level of aspiration situation can be found, nevertheless, for clinical evaluation the method has many advantages. One such advantage is the opportunity for the direct study of behavior as contrasted with projective methods which involve an additional step of interpretation and consequently a potential, additional source of error.

This is a careful and insightful work, well worth the study of psychologists interested in the utilization of level of aspiration techniques for the clinical or experimental study of personality. Its many fruitful hypotheses should serve as a stimulation for much needed additional research in this field.

<div align="right">Julian B. Rotter</div>

Ohio State University

Goldstein, Naomi F. *The roots of prejudice against the Negro in the United States.* Boston: Boston Univ. Press, 1948. Pp. ix+213. $2.50.

In this book, Dr. Goldstein advances the thesis that present attitudes of prejudice against the Negro in the United States can only be understood in terms of the unique position of the Negro as a former member of a slave class. In developing this point of view, Dr. Goldstein uses an approach that is becoming more and more prevalent among well-trained students of the social sciences. The materials and methods of analysis of history, economics, and sociology, as well as psychology, are brought to bear upon the problem with the happy consequence that this small volume gives a well-rounded picture of the many facets of race prejudice.

The book begins with a brief description of the status of the Negro in the United States today. The concept of race prejudice is then developed as a tendency to react to an individual not primarily as an individual but as a member of a racial group. "In examining a situation to determine the presence or absence of race prejudice, the main criterion is reaction to the group rather than to the individual, any attitude at all which meets this definition must be considered prejudiced" (pp. 34–35). To be free from race prejudice does not, of course, imply that one must ignore any or all "unfavorable" facts relating to the members of a racial group, but rather that these facts, if they exist, must be evaluated in terms of some frame of reference other than racial.

A critical examination of existing theories of race prejudice is then undertaken with the result that the author concludes that no single theory can satisfactorily account for the continued existence of prejudice against the Negro in the United States. Current prejudice can be traced to the institution of slavery, the maintenance of which demanded legal and psychological separation of white and Negro. The traditions, beliefs, customs, and attitudes established at the time of slavery became crystallized in social norms governing Negro and white relations. With the emancipation of the Negro, slavery was no longer legally permissible, but the norms derived from slavery were still in existence and were perpetuated by the legalization of discriminatory and segregative practices. Prejudice and hostility thus continue to exist because of the character of legally supported segregation and discrimination, dating from the period of Reconstruction.

Of interest to many psychologists will be Dr. Goldstein's analysis of the way in which the norms concerning the Negro are expressed in songs, jokes, cartoons, newspaper stories, drama, films, fiction, radio—and even advertisements. This analysis reveals three basic stereotypes: the picture of the Negro as a "contented slave," the picture of the Negro as a "brute barbarian," and the picture of the Negro as a "comic." The latter has been the most pervasive of the stereotypes and embodies all of the beliefs as to why the Negro was happy as a slave and wretched as a freeman—his love of fun, his dependency upon white folks, his childish thinking, his inherent laziness, his few and simple needs, and so forth. These stereotypes have, at all times, served to justify the status of the Negro as a second-class citizen.

The development and perpetuation of a norm which is essentially hostile rather than friendly toward the Negro, Dr. Goldstein believes, is the result of a system of punishments applied to those who refuse to conform to the norm. For a white person to engage in friendly relations with a Negro results in loss of prestige, status, and other social and economic rewards. "Prejudice against the Negro cannot reasonably be expected to disappear until segregation—the *forced isolation* of all members of one group from all members of the other is no longer legally defensible" (p. 120).

Allen L. Edwards

The University of Washington

Kaufmann, Fritz. *Your job.* New York: Harper & Brothers, 1948. Pp. xii + 238. $2.75.

Your job, a guide to opportunity and security, is a factual and commonsense treatment of some pertinent facts for the poorly informed individual

in the labor market. Although it is optimistically "written for every worker" and "lay and professional adviser," it is too elementary to benefit such a wide audience. However, it can definitely be read with profit by most entry workers, job explorers, and less experienced counselors and personnel workers. The book might also serve as a supplementary text in college counseling and guidance courses.

Early chapters are devoted to a discussion of elementary self-analysis and a survey of the world of work, while later chapters deal with such topics as wages, personal documents and papers, where to go for information about jobs and job opportunities, and the job interview. There are also helpful sections on training and schooling, the role of labor unions, setting up your own business, and rights and benefits under current social legislation.

Your job is recommended as a book intended for the inexperienced and poorly informed worker and for the fledgling personnel administrator or counselor. There is a little uneveness in coverage of some topics and perhaps several places where we might take exception to the author's emphasis, yet these minor criticisms are more than balanced by the book's merits for this specified audience. It covers crucial aspects of "your job" in an interesting and readable manner. It presents much information that the counselee could profitably read in preparation for a counseling interview. It is based on excellent source material (largely federal and local government publications). It is a realistic discussion of jobs with frank statements about common employer attitudes and suggestions for handling some of these prejudices. It does not overlook the tremendous importance of individual counseling.

As an introductory source book and orientation to the topic of the individual and his job, Kaufmann's book is commendable.

William A. McClelland

Brown University

New Books, Monographs, and Pamphlets

Selecting the new employee. Paul W. Boynton. New York: Harper and Brothers, 1949. Pp. 136. $2.00.

Personal adjustment in old age. Ruth S. Cavan et al. Chicago: Science Research Associates, Inc., 1949. Pp. 204. $2.95.

Readings in general psychology. Wayne Dennis, Editor. New York: Prentice-Hall, Inc., 1949. Pp. 525. $3.75.

An analytical bibliography of modern language teaching. Edited by Robert Herndon Fife. Washington, D. C.: American Council on Education, 1949. Pp. 549. $5.50.

The validity of commonly employed occupational tests. Edwin E. Ghiselli. Los Angeles: University of California Press, 1949. Pp. 287. $.75.

Psychology for the profession of nursing. Jeanne G. Gilbert and Robert D. Weitz, New York: The Ronald Press Co., 1949. Pp. 275. $3.00.

Man in the primitive world. E. Adamson Hoebel. New York: McGraw-Hill Book Co. Inc., 1949. Pp. 543. $5.00.

Experiments on mass communication. Carl I. Hovland, Arthur A. Lumsdaine, and Fred D. Sheffield. Princeton: Princeton University Press, 1949. Pp. 345. $5.00.

Directed thinking. George Humphrey. New York: Dodd, Mead and Co., 1949. Pp. 229. $3.50.

These are your children. Gladys G. Jenkins, Helen Shacter, and William W. Bauer. Chicago: Scott, Foresman and Co., 1949. Pp. 192. $3.50.

The nature and conditions of learning. Howard L. Kingsley. New York: Prentice-Hall, Inc., 1949. Pp. 579. $4.50.

Frustration. The study of behavior without a goal. Norman R. F. Maier. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 264. $3.50.

Guidance policy and practice. Robert H. Mathewson. New York: Harper and Brothers, 1949. Pp. 294. $3.00.

The psychology of personal adjustment. Second edition. Fred McKinney. New York: John Wiley and Sons, Inc., 1949. Pp. 752. $6.00.

Polls and public opinion. Norman C. Meier and Harold W. Saunders. New York: Henry Holt and Co., 1949. Pp. 400. $3.50 College edition; $4.75 Trade edition.

Psychological testing. James L. Mursell. Second edition. New York: Longmans, Green and Co., 1949. Pp. 488. $4.00.

A manual of pronunciation. Norriss H. Needleman. New York: Barnes and Noble, Inc., 1949. Pp. 323. $4.00.

Child development. Willard C. Olson. Boston: D. C. Heath and Co., 1949. Pp. 432. $4.00.

Biology of mental defect. Lionel S. Penrose. New York: Grune and Stratton, Inc., 1949. Pp. 270. $4.75.

Christianity and fear. Oscar Pfister. New York: The Macmillan Co., 1949. Pp. 589. $6.50.

Effective communication in industry. Paul Pigors. New York: National Association of Manufacturers, 1949. Pp. 88. Copies free upon request.

Experimental psychology. Leo Postman and James P. Egan. New York: Harper and Brothers, 1949. Pp. 500. $4.50.

Opportunities in vocational guidance. Sarah Splaver. New York: Vocational Guidance Manuals, 1949. Pp. 104. $1.00.

Children of Brasstown. Celia Burns Stendler. Urbana: Bureau of Research and Service, College of Education, University of Illinois, 1949. Pp. 103. $.60.

Introduction to Zen Buddhism. Daisetz Teitaro Suzuki. New York: Philosophical Library, 1949. Pp. 136. $3.75.

Adolescent fantasy. Percival M. Symonds. New York: Columbia University Press, 1949. Pp. 397. $6.00.

Experimental psychology. Benton J. Underwood. New York: Appleton-Century-Crofts, Inc., 1949. Pp. 638. $4.50.

Children with mental and physical handicaps. J. E. Wallace Wallin. New York: Prentice-Hall, Inc., 1949. Pp. 576. $5.00.

Advances in insurance coverage—accident prevention and control. New York: American Management Association, 1948. Pp. 39. $.75.

Appraising and training office supervisors. New York: American Management Association, 1948. Pp. 39. $.75.

Building quality into manpower. New York: American Management Association, 1948. Pp. 35. $.75.

Study of Executive Leadership in Business.
I. The R, A, and D Scales

C. G. Browne*
Wayne University

This is the first in a series of papers which will present the following methods for the study of leadership and executive relationships in business: R, A, and D scales; social and organizational contacts; sociometric pattern; Goal and Achievement index (1).

The total study proceeded on the following hypotheses: (1) leadership is a process based upon the inter-relationships of individuals in a group which is working toward a goal that has been accepted as desirable; (2) executive function and leadership in business is a process of the inter-action of social and working relationships within and outside of the executive groups; and (3) executive and leader relationships can be analyzed through the application of methods which are not designed to measure personal executive traits as psychological entities.

Procedure

The subjects in these explorations were 24 executives of a tire and rubber company in Ohio, named the Congo Tire and Rubber Company for purposes of the study. Table 1 includes a listing of the executives by title and department. All of the company executives on the first, second, and fourth echelons of the business, and all of the executives on the third echelon with one exception were included. Data were obtained in a moderately structured interview, varying in length from 2½ to 3½ hours. Some of the executives completed the R, A, and D scales during the interview, while others completed them at another time. In all cases, the scales were explained during the interview.

R, A, and D Scales

The RAD index form devised by Stogdill and Shartle in their studies of Naval leadership consists of six scales, each containing eight state-

---

ments (4). Scales A and B are for Responsibility; scales C and D, for Authority; and scales E and F, for Delegation of authority. The person completing the forms checks his first and second choices of statements as they best apply to him on each of the six scales.[1] The following are examples of the statements for each of the variables: *Responsibility,* "I am responsible for the successful operation and coordination of all activities in the organization"; *Authority,* "I make no decisions whatsoever but request instructions from my superior on all matters"; *Delegation of authority,* "I have delegated full authority to my assistants, allowing them complete right of decision in all functions."

Scoring of the individual items on each scale was developed using the Thurstone equal appearing interval technique (5). To establish scale values, the statements were evaluated by staff, graduate students, and seniors in psychology at the Ohio State University. The mean of the point values of the four statements checked is the score on that variable. Scale values for the statements range from 1.0 (indicating a high degree of the factor) to 8.7 (indicating a low degree). *Therefore, the lower scores indicate a higher degree of the item measured, while the higher scores indicate a lower degree.*

R, A, and D Scores

The R, A, and D scores of each executive, the mean scores by departmental and total groups, and the range for each factor are given in Table 1. Remembering that the lower scores indicate a higher degree of the factor, the score of 1.6 for the president and general manager represents the highest for both R and A. Likewise, the scores of 5.2 and 5.1 for the manager of tube sales represent the lowest for R and A, respectively. The vice-president-sales had the highest D score, and he was also one of the three executives who received the greatest number of choices on the sociometric diagram. While the secretary of the company had the lowest D score (6.7), an analysis of his work revealed that he had no one under his supervision to whom he could delegate the relatively small degree of authority which he estimated he had.

In any measure—individual scores, departmental means, or total means—the R scores were almost consistently the highest, followed by the A scores, and finally the D scores. This indicates a general trend for the executives to estimate that they delegated authority in a lesser amount than they estimated either their responsibility or authority, and that their authority was less than their responsibility. Although the ranges of the R and A scores were almost identical, there was a concentration of R scores, there being 22 cases between 2.7 and 3.9, with a mode of

Table 1

R, A, and D Scores

Executive Department and Title	R Score	A Score	D Score
General Administration			
President and genl. manager	1.6	1.6	2.7
Secretary of the company	3.2	4.3	6.7
Director public relations	3.3	2.9	3.8
Purchasing agent	3.9	3.6	6.4
Department mean	3.0	3.1	4.9
Sales			
Vice-president-sales	2.7	3.5	2.3
Sales manager	2.7	2.9	3.0
Manager Congo stores	2.7	3.8	4.5
Manager sales promotion	2.7	3.4	4.9
Manager sales orders	2.7	4.4	4.9
Manager tube sales	5.2	5.1	5.1
Department mean	3.1	3.8	4.1
Finance			
Treasurer	2.7	4.1	4.9
Comptroller	2.7	3.4	3.8
Supervisor cost accounting	2.9	4.1	5.2
Chief accountant	3.3	3.4	6.2
Department mean	2.9	3.7	5.0
Manufacturing			
Vice-president-manufacturing	3.0	2.9	3.2
Plant engineer	3.7	4.4	4.0
Chief chemist	3.7	3.4	3.8
Product engineer	2.7	3.6	3.8
Foreman bicycle tire production	3.7	4.7	2.9
Manager production control	2.9	3.4	4.4
Manager quality control	2.7	4.5	5.5
Manager shipping	3.3	4.3	6.0
Department mean	3.2	3.9	4.2
Personnel			
Personnel director	3.2	2.3	2.5
Industrial engineer	3.2	2.3	3.3
Department mean	3.2	2.3	2.9
Total group mean	3.1	3.6	4.3
Range	1.6–5.2	1.6–5.1	2.3–6.7

2.8 for all R scores. The A scores, however, distributed more uniformly, with a mode of 4.3, while the D scores had the greatest range, but were distributed most uniformly. There were 17 D scores lower than 3.3, compared with 18 R scores of 3.3 or higher.

While these scores cannot be considered to be predictive of executives in other companies on the basis of work done, or departmental assignment, or echelon level, the method offers opportunities to study working

relationships between executives which may be related to any of these variables. As a measure of communication within the company and of other personal relationships, the R, A, and D scales offer further possibilities. These measures might be obtained by a study in which an executive's seniors complete R and A forms for him and his juniors complete D forms for him. A comparison of these scores with the executive's own forms would constitute a measure of the individual's understanding of his responsibility and authority from the seniors who determine them and of his delegation of authority from the juniors to whom delegation is made.

R, A, and D Relationships

Table 2 includes correlation coefficients between R, A, and D and other variables used in the study. *These correlations are descriptive only of the*

Table 2

Product Moment Inter-correlations of R, A, and D Scores
and R, A, and D Scores Correlated with Other Variables

Variable (N = 24)	R	A	D
R (Responsibility)	x	.56	.29
A (Authority)	.56	x	.54
D (Delegation of authority)	.29	.54	x
Time spent in supervision**	−.06*	−.25*	−.12*
Number of choices***	.29*	.28*	.48*
Executive's salary	.48*	.41*	.49*
Executive's echelon	.34	.40	.14

* The sign for this correlation has been changed so that in interpreting the correlations a large score in one variable is also indicative of a large score in or a greater degree of the second variable.

** This variable was expressed in the per cent of the executive's total time which he estimated he spent in supervision.

*** As determined from the sociometric diagram.

relationships existing between the variables for this particular population of executives. They cannot be interpreted as sampling statistics, since the group of executives studied here does not constitute a statistical sample.

The inter-correlations between the three factors were .56 for R and A; .29 for R and D; and .54 for A and D. In the studies of Naval leadership, unpublished correlations for a group of 40 Naval officers were found to be .56 for R and A; .16 for R and D; and .86 for A and D. These comparative correlations between the business executives and the Naval officers indicate the same general trend in the inter-correlations between factors, although the Navy correlation of .86 for A and D was considerably higher than the executive correlation of .54 for the same factors.

Some of this difference may have been due to the possibility that such concepts as authority and the delegation of it are more clearly defined for military personnel than they are for business executives, and that they are measured and weighed with greater absoluteness in the military environment.

The correlation between authority and time spent in supervision, the largest of the three negative correlations between these variables, was −.25. This indicates that the executive who devoted a greater percentage of his time to supervisory activities, as contrasted with such other activities as planning or coordination or evaluation, tended to have an A score which represented a lesser amount of authority.

In a later paper, the sociometric pattern which was used in the study will be discussed. The "number of choices" variable was determined from the listing which each executive made of the men with whom he spent most time in getting his work done. In the sociometric diagram, the greatest number of choices was received by three executives who were in the second echelon. The relatively high positive correlation of .48 between number of choices and D score indicates that those men who were consulted most and with whom most time was spent in getting work done also tended to be the men who were delegating authority to the greatest degree.

The correlations between R, A, and D scores and salary all indicate that executives with the higher salaries tended to have scores which indicated a greater degree of the three factors. In view of the low correlation between echelon and D score, the relatively high correlation of .49 between salary and D score may be surprising. However, this correlation is strongly influenced by the fact that several of the executives on the fourth echelon were receiving higher salaries than some of the executives on the second and third echelon.

The correlations between R, A, and D scores and echelon were not as high as they were with salary. However, it is quite logical that the correlations with echelon were highest for responsibility and authority, since it can be expected that the higher level executives would have a higher index on these factors. Delegation of authority, on the other hand, is an individualized factor, not greatly related to the executive's echelon.

Summary

The R, A, and D scales introduced by Stogdill and Shartle in their studies of Naval leadership were applied to a group of 24 executives in a tire and rubber manufacturing company. The scores for each executive on each of the three factors provide a measure of the individual's evaluation of his responsibility, authority, and delegation of authority. From

C. G. Browne

their scores, these executives estimated that their responsibility and authority were greater than their delegation of authority.

Since the factors measured by the R, A, and D scales are particularly important at the executive level, a quantitative method such as presented here should aid in the analysis and understanding of executive functioning and business leadership. This is based upon the general hypothesis that leadership and executive activity are dependent upon social and working relationships in group activities, and that their study from this approach will prove more helpful than the analysis of individual characteristics with psychological trait testing methods has proven.

References

1. Browne, C. G. *An exploration into the use of certain methods for the study of executive function in business.* Unpublished Ph.D. dissertation, The Ohio State University, 1948.
2. Jenkins, W. O. A review of leadership studies with particular reference to military problems. *Psychol. Bull.*, 1947, 44, 54–79.
3. Stogdill, R. M. Personal factors associated with leadership—a survey of the literature. *J. Psychol.*, 1948, 25, 35–71.
4. Stogdill, R. M., and Shartle, C. L. Methods for determining patterns of leadership behavior in relation to organization structure and objectives. *J. appl. Psychol.*, 1948, 32, 286–291.
5. Thurstone, L. L., and Chave, E. J. *The measurement of attitude.* Chicago: Univ. of Chicago Press, 1929.

An Analysis of the Leaderless Group Discussion

Bernard M. Bass

Louisiana State University

Several studies in recent years have reported the use of the leaderless group discussion situation as an aid in selecting candidates for positions involving leadership (1, 2, 3, 4). Little has been done to evaluate this technique quantitatively or to investigate the possibility of making objective measures of individuals in this situation. The purposes of the present study were to investigate the extent of agreement among raters of discussion participants and the relation between the total amount of time a participant spent talking[1] in the leaderless group discussion, and the ratings he obtained.

Subjects, Method and Apparatus

A class of 20 educational psychology students served as subjects. Twelve were men and 8 were women. They ranged from freshmen to graduate student, with a median of 2 years college education. Several students had 2 or more years of teaching experience.

A total of 6 leaderless group discussions was run in 6 weeks. The 20 subjects were divided randomly into Group A and Group B. Group A participated in the first discussion while Group B observed the participants. The two groups of 10 students each switched roles for the second discussion. Those 10 participants of the first 2 discussions who had been given the highest leadership ratings for the discussions by their classmates formed the third leaderless group discussion. The fourth discussion was composed of the remaining 10 participants who had received the lowest leadership ratings. The original groups, A and B, were used again for the last 2 discussions. Group A participated in and Group B observed the fifth discussion, and Group B participated in and Group A observed the sixth discussion.

Each of the discussions lasted 20 minutes and was held during class hours in the classroom. The 10 participants were seated around the outside of a V-shaped table. A code numbered place card was put in front of each of the participants for identification. The 10 observers sat facing the participants on the other side of the room. To provide adequate

[1] For a discussion of the use of time spent talking in the individual interview as a predictive measure, the reader may refer to Chapple, E. D. and Donald, G. A method of evaluating supervisory personnel. *Harv. Bus. Rev.*, 1946, 24, 197-214.

motivation for the participants, each of the discussions were considered a course examination and grades from A to E were awarded by the experimenter-instructor.

Oral instructions were as follows:

"You will be given a problem and will have thirty minutes in which to discuss it. You will be graded, not only on how well you as an individual contribute to the group discussion, but also on how well the group does as a whole.

"Everyone may receive an A or everyone may receive an E depending on how much he contributes and how much the group progresses. Therefore, if you feel someone else is 'off the track,' is wasting the group's time and therefore is lowering your grade, feel free to cut in and get the group back on its proper assignment."

The problem presented to a discussion group was pertinent to the material on educational psychology which the subjects had supposedly studied the night before. It called for the development of a program or set of plans which must be sold to another group, such as a school board. The problem was read aloud to the group twice and then they were told to begin discussing it.

One example of the problems used is as follows:

The School Board of your town has gone progressive. The Board realizes that teachers cannot do everything and are planning to obtain a staff of specialists in various areas to cope with the several problems which teachers are unable to handle adequately. Consider yourselves as the chairmen of the ten departments of your high school of 5,000 students. You are meeting thirty minutes before the School Board goes into session. The present high school personnel consists of teachers, the principal, an office staff, and a janitorial staff. Your problem is to agree upon the four specialists you will ask for, and the reasons you will present for choosing those four. The School Board will only appropriate $12,000. And remember, there are 5,000 students, so don't plan on overloading the four specialists.

For the first 3 discussions, the experimenter used a stop-watch to clock the number of seconds each participant talked and then recorded the measurement under the participant's name on a log sheet.[2] It was found difficult to keep up with the constant shift in speakers, especially during arguments. Interruptions and pauses within speeches also increased the difficulty of measuring and recording. Fortunately, the differences among participants were large enough to allow one to assume that the measurement errors were not of such an order as to warrant discarding the results of the first 3 discussions.

In order to increase the time data precision, the Group Discussion Chronometer was designed by the investigator and introduced into the study in the fourth discussion. The GDC consisted of a panel of 10

[2] Wire recordings were not used because of the difficulties of dubbing in the speakers' code numbers which would have been necessary in order to later identify who was speaking at a given time.

button switches spaced to allow each of the experimenter's fingers to operate 1 button without having to move his hands. Each push-button closed the circuit of 1 of 10 sweep-hand, self-starting, electric clocks mounted on a board outside of the classroom and connected by a cable to the switch panel. One button and 1 clock were devoted to each participant and a cumulative measure, in seconds, was obtained of the total time each participant talked. The experimenter pressed the button appropriate to the participant each time the participant began talking and released it when the speaker stopped or paused. Unlike the stop-watch, the GDC could record individual times even when two participants attempted to talk at the same time.

Both participants and observers rated participants by 3 different rating methods on 13 items. The 3 rating methods were the "Spread N," which was used for all 6 discussions; the "3 Best minus 3 Worst," which was used for the third discussion, and the "rank order of merit," which was used for the fourth discussion. All ratings were recorded on a prepared inventory by the raters immediately after the conclusion of a discussion.

Where 10 participants were to be rated, the "Spread N" technique[3] was as follows:

Each of the 20 observers and participants assigned 10 votes on a given item to the participants in any way desired. If the rater thought one person completely outclassed all the other 9 on a given attribute, the rater assigned all 10 votes to the 1 individual and none to the others. The rater gave 1 vote to each of the 10 participants if all participants were judged equal. Any other distribution of the 10 votes was possible.

When the rank order of merit method was used, the raters were instructed to rank the 10 participants on each of the 13 items. For the "3 Best minus 3 Worst" method, the raters were told to select the 3 highest and 3 lowest participants on each of the items.

The 13 items upon which appraisals were made were as follows:

Vote for the person or persons:

1. Whom you think led the discussion.
2. Whom you think knew most about the topic discussed.
3. Whom you think most influenced the other participants in the discussion.
4. Who most clearly defined the problems, who brought them into sharp focus, and who best organized the group's thinking during the discussion.
5. Whom you would select to be superintendent of schools if he (she or they) had the proper experience and training.
6. Whom you like best.
7. Who offered the best solutions to the problems discussed.

---

[3] The reader can recognize a similarity between this technique and the pooled judgement method for differentially weighting traits of a composite criterion described in Burtt, H. E. *Principles of employment psychology.* New York: Harper, 1942, 354.

8. Whom you would like to see as chairman or head of the department.
9. Who most motivated the others to participate in the discussion.
10. Who seemed most interested in the discussion.
11. Whose class you would like most to be in, if all the participants were teachers.
12. Whom you would select to address an audience of teachers.
13. Who should get the best grade for today's discussion.

For the "Spread N" rating procedure, the total number of votes received from all those who rated a participant were divided by the number of raters to obtain his mean rating. To obtain an individual's rating by means of the "3 Best minus 3 Worst" technique, the number of "worst" votes received from all the raters were subtracted from the number of "best" votes. For the ranking procedure, a participant's average rank was computed. As will be shown later, there seemed little value in computing mean ratings for each of the thirteen items, item by item, as they all seemed to be measuring the same variable—leadership status.

Agreement Among Raters

It was felt unprofitable to compute the 190 intercorrelations between raters because of the small number of ratees. Participants' total "Spread N" ratings in the first, second, fifth, and sixth discussions assigned by a given rater, were converted instead into ranks, and the average intercorrelation of all rank orders was computed to give a rough appraisal of the extent of agreement among judges (See Woodworth (5), p. 372 ff.). The average intercorrelations obtained for the first and last two discussions were .72, .61, .63, and .41 respectively.

The correlation between combined participant's ratings and combined observer's ratings of each participant was another measure of the extent of agreement among raters. When participants' ratings for the first 2

Table 1

The Rate-Rerate Reliabilities of Nineteen Judges Rating Participants in Discussions I and II and Six Weeks Later in Discussions V and VI

Judge	r	Judge	r	Judge	r
1	.66	8	.65	15	.63
2	.64	9	.53	16	.86
3	.90	10	.86	17	.84
4	.48	11	.69	18	.41
5	.79	12	.72	19	.53
6	.76	13	.76	20	.72
7	*	14	.67		

* Judge 7 failed to attend the last two discussions.

discussions were combined into 1 distribution, there was a correlation[4] of .90 between ratings assigned to participants, by participants, and by observers.

Retest and Rerate Reliability

Each judge's ratings of participants of the first 2 discussions were correlated with the ratings assigned by the judges to the same individuals when they acted as participants of the last 2 discussions 6 weeks later. The fifth discussion was among the same participants as the first, and the sixth discussion was among the same participants as the second. Table 1 lists the 19 rate-rerate reliability coefficients obtained. By converting these coefficients into Fisher's Z-function, a mean coefficient of .72 was obtained. There was a correlation of .87 between the total time participants talked in the first or second discussion, and the time they talked in the fifth or sixth discussion.

The 10 subjects of the 20 who participated in the first 2 discussions who received the highest ratings were placed in the third discussion. The other 10 subjects were placed in the fourth discussion. Despite some reversals, those in the highest leadership status in discussions I or II became the "leaders of leaders" in discussion III and the highest "followers" in discussions I or II became the "leaders of the followers" in discussion IV. Rank order correlations between participants' ratings in discussions I or II and III or IV were .78 and .76 respectively. The conclusion may be reached that despite the change in groups and restriction of range, leadership status tended to be generalized from one leaderless group discussion to another. There seemed to be consistency in both the behavior of participants and the ratings they received in the leaderless group discussion.

Interrelationships Among Variables

The number of votes assigned to participants of discussion I or II by all raters were combined, item by item, and correlated with the time participants spent talking. Table 2 shows the results obtained. When votes obtained on all 13 items were correlated with time spent talking, the coefficient obtained was .93. Because of the high correlations obtained between each item and time spent talking, it was thought unnecessary to compute the matrix of intercorrelations, as all seemed to be measuring the same factor, i.e. leadership status. The unidimensionality may have been due in part to halo effect.

Discussion V and VI showed similar results. The correlation between time and leadership status ratings was .86. Taking into account the high rate-rerate reliabilities mentioned previously, it seems that time

[4] Unless otherwise stated, correlations were obtained by the Pearson product-moment formula.

Table 2

Correlations Between Time Spent Talking in the First Two Discussions and
Number of Votes Received on Each of Thirteen Leadership Items

Item	r	Item	r	Item	r
1	.91	6	.85	11	.86
2	.87	7	.92	12	.87
3	.90	8	.82	13	.92
4	.89	9	.90	All Items	
5	.85	10	.87	Combined	.93

spent talking and leadership status are closely associated and that this close association may be generalized from one leaderless group discussion to another.

Comparison Among Rating Methods

The effects of using a different rating technique were negligible. In discussion III there was an almost perfect linear relationship between "Spread N" ratings assigned by 20 raters, and the "3 Best minus 3 Worst" ratings. In discussion IV, an almost perfect curvilinear relationship of the order $y = a + b \log X$ was found to exist between ranking and the "Spread N" ratings, but whereas the former tended to distribute individuals evenly, the "Spread N" like the "3 Best minus 3 Worst" method tended to dichotomize participants. The "Spread N" seemed to scatter out the leaders more widely, while the "3 Best minus 3 Worst" best spread out the followers. From a realistic approach, on the basis of this extremely small sample, the "Spread N" would seem to be the most valuable as a technique for selecting from the top end of a distribution. Leaders were distributed by the "Spread N" in the same manner as on the variable of time spent talking. The correlation between leadership status and time spent talking was therefore highest with this rating method. Since the leaderless group discussion was designed primarily to discriminate among leaders, the "Spread N" best approaches the needs of the situation and the correlation obtained when using this rating method, does not appear to be spurious to any great extent.

Interpretation and Conclusions

Several hypotheses can be drawn for further investigation from the results obtained and from personal observations of the leaderless group discussion.

1. If a group is given a verbal problem, with suitable motivation to cooperate and achieve the goals relevant to the problem, a differentiation of function will occur within the group.

2. In a leaderless group discussion, one task may be assumed by several people; some tasks may be assumed by one; some tasks may not be performed at all. These tasks include initiation or formulation of the problems and goals, organization of the group's thinking, clarifying other individuals' responses, integrating responses of several individuals, questioning, motivating others to respond, accepting or rejecting other individuals' responses, outlining the discussion, summarizing, generalizing, obtaining the group's agreement and formulating conclusions.

3. Because of the verbal nature of the situation, the more tasks an individual assumes, the more time he is forced to spend in talking to the rest of the group.

4. It is assumed that those individuals who carry out the above-mentioned tasks are perceived by others to be the leaders of the group discussion.

5. If the above hypotheses are correct, then the time an individual spends in talking in the leaderless group discussion is indicative of his status as a leader or follower in that group situation.

References

1. Brody, W. Judging candidates by observing them in unsupervised group discussion. *Personnel J.*, 1947 26, 170–173.
2. Frazer, J. M. New-type selection boards in industry. *Occup. Psychol.*, 1947, 11, 170–178.
3. OSS Assessment of Staff. *Assessment of men.* New York: Rinehart, 1948.
4. Taft, R. Use of the "group situation observation" method in the selection of trainee executives. *J. appl. Psychol.*, 1948, 32, 587–594.
5. Woodworth, R. S. *Experimental psychology.* New York: Holt, 1938.

Performance on the File-Remmers Test, How Supervise?
Before and After a Course in Psychology

Harry W. Karn

Carnegie Institute of Technology

A recently developed instrument for the measurement of the attitudes and understandings necessary for supervisory success is the File-Remmers questionnaire, *How Supervise?* The construction of the original forms of this test is described in an article by File (1). In a summary of the preliminary evidence on the validity of the device, File and Remmers (2) report significant increases in scores after supervisory training and significant differences in scores between successful supervisors and individuals by-passed because of lack of supervisory ability.

An indication that the test is lacking in universal validity is the study by Sartain (4). This investigator reports the test to have little or no predictive value for success in supervision in an aircraft factory. In an attempt to reconcile these findings with their positive evidence, File and Remmers (2) question whether Sartain's 40 supervisors in an expanded war industry are sufficiently typical to be used as cases for drawing general conclusions concerning the usefulness of tests in selecting supervisory personnel.

In view of the evidence reported to date, further studies appear to be in order before the question concerning the universal validity of the test can be settled. As a contribution toward this end, the following report is presented of an investigation designed to determine the effect of a psychology course upon college students' understanding of supervisory skills and practices as measured by the File-Remmers test.

Procedure

Forms A plus B of the test were administered under standard instructions to 108 students (104 males) in the College of Engineering and Science at Carnegie Institute of Technology during the first week of a first semester required course in psychology. About 98 per cent of this group consisted of students in their junior college year with the remainder consisting of irregulars in the sophomore and senior years. These students made up the *training group*. The same tests were administered to a comparable *control group* of 104 students (101 males) during the first week of a required course in English. During the last week of the semester both groups were again administered both forms of the test.

During the interim between testings the specific issues covered in the test were not discussed in either the psychology or English courses. Scores were not divulged nor the purpose of the investigation mentioned until after the final testings.

The training group consisted of five sections taught by four different instructors. This group took a three-credit course in general psychology. All of the instructors used standard psychological text books although the emphasis throughout the course was not upon text book content per se but upon the application of psychological principles to adjustment problems, particularly those likely to be encountered in human relations situations in industry. To this end, realistic case problems were dealt with and the student encouraged to solve them through the use of psychology and a systematic, analytical problem solving procedure.

The control group was made up of five sections of a three-credit course in English literature taught by five different instructors.

A feature of the investigation is the use of a control group with which to compare changes in test performance by the training or experimental group. This type of experimental design is mandatory if valid conclusions concerning the effectiveness of any training program are to be drawn. Changes in performance on the part of the experimental group can be attributed to the training program only if these changes are significantly different from any changes that may appear in the control group.

Results

Table 1 summarizes the essential statistical data for a comparison of scores made by the training and control groups on Forms A plus B of the test during the original and terminal testings. These data indicate a slight difference in mean raw scores between groups on the first test, an

Table 1

Comparison of Initial and Terminal
Scores on Forms A plus B for
Training and Control Groups

Group	N		Mean	S.D.	Critical Ratio*	r between Original and Terminal Scores
Training	108	Before	95.3	15.0		
		After	104.5	13.3	7.46	.61
Control	104	Before	97.1	16.9		
		After	98.9	17.9	1.70	.81

* Critical ratio of difference between groups on first test, .81. Critical ratio of difference between groups on terminal test, 2.54.

insignificant increase on the part of the control group on the retest and a highly significant increase of nine points of raw score on the retest by the training group. The nine point increase in mean raw score at this level is equivalent to a shift from about the 50th percentile to about the 70th percentile, according to the File-Remmer norms for Higher Level Supervisors (3). A comparison between retest scores of the training and control groups shows a difference of about five and a half points in mean score which is significant at nearly the one per cent level of confidence. This comparison, however, obscures the total gain made by the training group since this group had a lower mean (95.3) than the control group (97.1) at the time of the initial testing. The full extent of the gain made by the training group is evident in the ensuing comparative analysis which treats the data in terms of absolute differences in scores between testings.

A summary of the analysis made in terms of the differences between scores on the initial and terminal testings for training and control groups is presented in Table 2. These data reveal the average difference between the initial scores and the higher terminal scores for the training group to be significantly greater than a comparable measure for the control group.

Table 2

Comparison of Training and Control Groups in Terms of
Differences Between Initial and Terminal Testings

	Training	Control
N	108	104
Mean of Differences	9.3 (gain)	1.6 (gain)
S.D.	12.8	10.9
Critical Ratio	4.73	

Because of greater reliability, both forms of the test are recommended by its designers over the single form. It is reported, however, that the single form yields scores which are sufficiently reliable for gaining information about a group as a whole. It appeared worthwhile, therefore, to compare the double form analysis from the present study with the data from single forms. Scores on Form B from the initial testings were compared with scores made on the same form on the terminal testings. Form B was used because any possible practice effects from having taken another form of the test would be equated on both initial and terminal testings.

Table 3 is a summary of the data from the single form analysis. There is little difference between mean scores of the two groups on the initial test, an insignificant increase by the control group on the retest, and a

Table 3

Comparison of Initial and Terminal Scores on Form B for Both Groups

Group	N		Mean	S.D.	Critical Ratio*	r between Original and Terminal Scores
Training	108	Before	50.3	8.0		
		After	54.1	7.5	6.61	.61
Control	104	Before	50.8	10.3		
		After	51.5	10.6	.85	.68

* Critical ratio of difference between groups on first test, .43.   Critical ratio of difference between groups on terminal test, 2.04.

highly significant increase of four points of raw score on the retest by the training group.   The increase of four points at this level is equivalent to a shift from the 55th percentile to the 70th percentile, according to the File-Remmers norms (3).   The difference of nearly three points in mean score between groups on the retest is significant at about the five per cent level of confidence.

The full extent of the gains made by the training group on the single test form is revealed in the summary of the data presented in Table 4. These data show the average difference between the initial and terminal scores for the training group to be greater than the comparable measure for the control group at about a one per cent significance level.

Table 4

Single Form B Comparison of Training and
Control Groups in Terms of Differences
Between Initial and Terminal Testings

	Training	Control
N	108	104
Mean of Differences	4.1 (gain)	1.4 (gain)
S.D.	7.5	7.5
Critical Ratio		2.56

In general, the results of the single form analysis corroborate those from the analysis of data from both forms although the differences for the latter attain a higher level of statistical significance than those for the former.

Discussion

On the assumption that a psychology course designed to improve understanding of the principles of human behavior in industrial situations actually accomplishes this goal, the superior terminal questionnaire per-

formance of the students having taken this course would indicate that the instrument is measuring those skills having to do with an understanding of the principles of the successful management of human relationships In support of this conclusion is the absence of significant improvement in scores on the part of a comparable group of students tested at the same times but without having taken the psychology course between testings. Since good supervision presumably requires a knowledge of human relations skills, the present study can be interpreted as evidence for the validity of the File-Remmers test as a means of measuring one aspect of supervisory success.

The present investigation is a contribution towards the establishment of the universal validity of the test since it deals for the first time with college students under academic instruction. Previous studies, which have shown positive gains with training, have been concerned with the effects of specific industrial training programs among on-the-job supervisory personnel. The demonstration of improved scores under a variety of training conditions indicates that the responses are the result of the application of basic principles to the problems rather than the acquisition of specific answers to the questionnaire items.

High scores on a test of the type under discussion are no guarantee that the individuals making such scores will be good supervisors. There must be additional evidence from other sources that such individuals will put into practice the knowledge they possess. The argument for the use of the test is still good, however, for obviously individuals cannot put into practice knowledge that they do not have. The final answer to the question of measuring and predicting success in supervisory situations will probably take the form of a composite index based on test scores, biographical data and on-the-job ratings by qualified observers.

Summary

A group of 108 college students in their junior year were administered Forms A plus B of the File-Remmers questionnaire, *How Supervise?* before and after a course in psychology. The same tests were administered to a comparable control group of 104 students before and after a course in English literature.

An analysis of the data in terms of mean scores for both groups on initial and terminal testings and in terms of differences between scores on the two testings shows significant gains in favor of the psychology group. This is true for both the double form data, i.e., the comparison of scores on Forms A plus B on initial and terminal tests and the single form comparison of scores on Form B only.

References

1. File, Q. W.   The measurement of supervisory quality in industry.   *J. appl. Psychol.*, 1945, 29, 323–337.
2. File, Q. W., and Remmers, H. H.   Studies in supervisory evaluation.   *J. appl. Psychol.*, 1946, 30, 421–425.
3. File, Q. W., and Remmers, H. H.   *How Supervise?* (revised manual) New York: 1948, Psychological Corporation, pp. 8.
4. Sartain, A. Q.   Relation between scores on certain standard tests and supervisory success in an aircraft factory.   *J. appl. Psychol.*, 1946, 30, 328–339.

The Prediction of Accidents of Taxicab Drivers

Edwin E. Ghiselli and Clarence W. Brown

University of California

The long history of investigations concerned with the effectiveness of tests in the prediction of accident proneness among vehicle operators might lead one to suspect that a wealth of information exists on this subject. Examination of typical reviews of the literature, however, points up the fact that empirical evaluations are by no means extensive and are quite restricted with respect to types of tests, being almost wholly concerned with apparatus tests (2, 5, 7). There appears to be general agreement that several kinds of reaction time tests, particularly those involving more complex functions, are tho most useful. Tho value of tests of sensory acuity and perception is less certain. In the paper and pencil test field only intelligence tests have had more than a cursory examination. The validity of this type of predictor is rather low, with a validity coefficient of the order of about .15.

In many situations apparatus tests cannot be employed in the selection of vehicle operators. Both initial and maintenance costs often are too high. In certain cases applicants must be tested in large groups, and in others personnel capable of operating apparatus tests are not available. Indeed, in some instances sheer lack of space for a permanent set up of testing equipment is an obstacle. However one might view the desirability of apparatus tests as compared with the paper and pencil variety, financial, administrative, and physical limitations may be the deciding factor in the choice of the type of tests that can be utilized. This leads, then, to a need for a closer scrutiny of the data available concerning the effectiveness of paper and pencil tests.

The low validity of intelligence tests cited earlier can be considered to assume some significance if only the very few superior individuals in a large number of applicants are to be considered. The Personnel Research Section of the Adjutant General's Office have reported eight validity coefficients for intelligence tests and nine for mechanical principles tests relative to performance on various types of road tests (6). In the large majority of these studies the coefficients were based on more than one hundred cases. For intelligence tests the validity coefficients range from .03 to .33, with a median of .18, and for mechanical principles tests the range of validity coefficients is from .00 to .40, with a median of .20.

Since driving skill might be expected to be related to safety of operation these coefficients at least are suggestive that paper and pencil tests might be helpful in predicting accidents. The present authors, together with E. W. Minium, investigated the usefulness for street car motormen of a variety of tests, including some of the paper and pencil variety (4). The criterion in this study was accidents for an eight-month period. Paper and pencil tapping and dotting tests were found to be the most valid, surpassing even apparatus tests (sensory acuity, distance perception, and simple reaction time). A coefficient of the order of about .35 is descriptive of the tapping and dotting tests. Tests of mechanical principles, judgment of distance by perspective, and judgment of linear distances were found to be less useful. For them a validity coefficient of .15 is representative.

Data such as the foregoing, meager though they may be, at least are encouraging. Further information regarding the effectiveness of paper and pencil tests would be most helpful in indicating fruitful lines for further research. The results of a single investigation should not be considered definitive but certainly would be a helpful addition to knowledge in such a relatively unexplored area. It was with this intent that the present investigation of the effectiveness of various types of predictors of accidents of taxicab drivers was undertaken.

Description of Predictors

Eight different speed tests and an interest inventory which yielded four different scores were used. The first five of these tests were the paper and pencil tests employed by Ghiselli, Brown, and Minium with street car motormen (4). Following is a description of each predictor.

Dotting. This test consists of a series of circles one-eighth inch in diameter, connected by lines, and irregularly spaced. The subject is instructed to place one dot in each circle. No pretest practice is given and the time limit is one-half minute.

Tapping. In this test the subject is presented with a series of circles of one-half inch in diameter and is instructed to put three dots in each circle. Again, no pretest practice is permitted and the time limit is one-half minute.

Judgment of Distance. The intent of this test is to measure capacity to judge distances between objects utilizing as cues only perspective and interposition. Each item is a schematic representation of a table top on which are placed four equally sized cubes. The positions of the cubes and the angle of view are different in the different items. The task is to judge which of three cubes is nearest to a designated key cube. The time allowed for this test is eight minutes.

Distance Discrimination. Each item in this test consists of a square in which there are three test points and a reference point. The task is to decide which of the test points is closest to the reference point. Various distracting lines wind among the points. The time limit is three and one-half minutes.

Mechanical Principles. This test consists of a series of pictorially presented problems illustrative of various mechanical principles. Most of the items are

concerned with the movement of vehicles and the operation of levers. Only simple mechanical principles are involved. The time for this test is eight minutes.

Numerical Problems. This is an arithmetic test involving the making of change and the computation of fares when various rates and lengths of trip are given. Five and one-half minutes are allowed for this test.

Speed of Reactions I. This test is an attempt to put in paper and pencil form a complex reaction test such as the Viteles Motorman Test. Each item consists of a square in which different letters appear in various spatial arrangements. Depending on the specific letters given in an item and their spatial arrangement, the subject makes a mark in one of five spatially differentiated circles placed below the square. This test is preceded by detailed instructions concerning the rules together with some examples for practice. On each page of the test proper the rules are given so the subject has them immediately available for reference. The time for this test is four and one-half minutes.

Speed of Reaction II. This test is the same as Speed of Reactions I except that the subject can never refer to the rules once the test is begun. Four minutes are allowed.

Interest. The interest inventory has no time limit. It yields four separate scores and a total score which is simply the sum of the part scores. For each of the four scales there are 24 items. In each item the subject chooses between two different occupations. He is instructed to choose on the basis of interest and ignore such matters as pay, vacations, and the like. The first scale has to do with *Occupational Level*, and compares a job roughly at the semiskilled level with a higher level job. The correct answer is the lower level job. The second scale is concerned with *Outside Occupations*, the correct choice being a job which is done out of doors rather than indoors. The third scale attempts to measure interest in occupations which require dealing with the public, and is termed *Dealing with People*. The last scale, *Related Occupations*, compares jobs involving the operation of vehicles with other types of jobs.

In addition to these tests certain personal data concerning the drivers were available. This information consisted of age, years of formal education, years of previous experience driving taxicabs, and years of experience operating other types of vehicles, either commercially or in the armed services. Since frequently considerable reliance is put on these variables in hiring drivers they, too, were studied in relation to accidents.

Subjects

The subjects used in this investigation were 67 men who applied for work and were employed as drivers by a taxicab company during the same three-month period. All men took the tests prior to being hired, and to some extent their scores were taken into account in the decision regarding their employment. In the selection process a profile of test and inventory scores was plotted for each person and those individuals who showed marked deficiencies were rejected. Greatest emphasis was given to the numerical problems and speed of reactions tests and to the interest inventory scores. Approximately one out of five applicants were rejected on this basis. With the exception of the speed of reactions test all men took all tests. For the speed of reactions tests the number of cases is 57.

Criterion

Accidents are a notoriously unreliable index of human behavior (5). Furthermore, accident proneness is by no means a unique trait. Examination of the relationships among different types of accidents incurred by operators of public conveyances indicates that there is considerable specificity (1). The problem of setting up a criterion to measure safety of performance, then, is fraught with many difficulties. In the present investigation the situation was complicated by the fact that only the safety records for the first five weeks of employment were available. However, this period is particularly critical since it is during this time that supervisors' judgments concerning the drivers' skill are crystallized. Of the 67 drivers, 48 incurred no accidents during the first five weeks of employment, 17 were involved in one accident, and two men were involved in two accidents. With such a distribution of accidents it was impossible to determine validity relative to number of accidents. As a consequence the men were divided into two groups, the accident free men (48 cases) and the accident group (19 cases), and biserial coefficients of correlation were utilized as indices of validity.

Results

In Table 1 are given the validity coefficients for the various predictors using accidents of the drivers during their first five weeks of employment as the criterion. It is apparent from this table that with the exception of the dotting and tapping tests, and possibly the interest inventory, none of the validity coefficients can be considered particularly significant.

The validity coefficients of the first five tests, dotting, tapping, judgment of distance, distance discrimination, and mechanical principles, are approximately of the order found earlier for motormen (4). The pre-

Table 1

Validity Coefficients for Various Predictors in Relation to Accidents

Validity Coefficient	Predictor	Validity Coefficient	Predictor
.35	Dotting	.23	Occupational Level (Interest)
.47	Tapping	.23	Outside Occupations (Interest)
.18	Judgment of Distance	.11	Dealing with People (Interest)
.20	Distance Discrimination	.20	Related Occupations (Interest)
.11	Mechanical Principles	−.08	Age
−.09	Numerical Problems	−.10	Years of Formal Education
−.04	Speed of Reactions I	.00	Years of Experience Driving Taxicabs
.00	Speed of Reactions II		
.28	Total Interest Inventory	.07	Years of Commercial and Service Driving

dictive power of the first two tests is fairly substantial while that of the latter three is low. On the basis of the motorman study weights were developed for these five tests for computation of a battery score. The effective weights are as follows: dotting, and tapping each four, judgment of distance and distance discrimination each two, and mechanical principles one. Applying these effective weights to the scores of the 67 taxicab drivers the validity coefficient of the battery was found to be .69. None of the accident group fell in the upper 25% of scores. Undoubtedly this coefficient is fortuitously high and would not be obtained with another similar sample. For the motormen the validity of this battery was of the order of .35. A battery of this type certainly seems worthy of further study for the selection of operators of vehicles. It is apparent that the tapping and dotting items are the most important components of the battery. Remembering that the total testing time for these two tests is only one minute the validity coefficients are surprisingly high. They are, in fact, too high to be readily acceptable just on the basis of two investigations, and further evaluation of these tests is indicated.

The fact that scores on the numerical problems test were found to be unrelated to accidents is not unexpected. There is no reason to suppose that ability to solve arithmetical problems is related to capacity to drive a motor vehicle safely. However, the complete lack of validity on the part of the speed of reactions tests certainly was not anticipated. At least, superficially, these tests seem to measure very nearly the same abilities as those measured by complex reaction tests, such as the Viteles Motorman Test, which have considerable predictive power. Perhaps these findings may be taken to indicate that certain kinds of abilities measured by apparatus tests and important in safe performance cannot be measured by paper and pencil prototypes.

The interest inventory that was utilized was designed principally to indicate which applicants would tend to stay on the job as taxicab drivers and which would tend to leave. In the area where this study was made labor turnover in the taxicab industry was quite high. An analysis of the situation indicated that a large proportion of the persons who left employment did so because they found that they did not like the nature of the work or at least they preferred other types of jobs. Nevertheless, the validity coefficient of .28 is suggestive of the value of interest measures in the prediction of accidents. The particular inventory utilized in this investigation undoubtedly can be greatly improved and the fact that three of the scales (occupational level, outside occupations, and related occupations) yielded validity coefficients of .20 cannot be ignored.

In the hiring of workers considerable emphasis is given to the factors of age, education, and work history. The subjects in the present investigation differed markedly in these variables. The range in each

variable is as follows: for age, 21 to 53 years, for education, seventh grade to college graduate, for previous taxicab experience none to 6 years, and for all types of experience driving vehicles other than private cars none to 23 years. In spite of the wide range of individual differences none of these factors was found to be appreciably related to safety of operation. In view of the success with which personal data has been used in the selection of employees in other occupations the relationships here may well be subject to some doubt. It is quite possible that the relationships are curvilinear and thus would not be adequately measured by biserial coefficients. The numbers of cases in the study was considered too small to warrant any detailed study of curvilinear relationships. However, inspection of the tabulations suggested an optimal age of 30 years and an optimal educational level of 10 grades for safe performance on the job. No similar optimums could be noted for previous driving experience.

Discussion

Taken together with the findings of previous investigations, the results of this study indicate that accidents can be predicted by paper and pencil tests. No brief can be made for the position that such tests invariably will be effective, but at least it can be said that they can be as effective as apparatus tests. It is likely, of course, that a combination of paper and pencil and apparatus tests would give better results than either the one or the other type alone. However, it may be questioned whether for purposes of employee selection the inclusion of apparatus tests increases the goodness of prediction sufficient to warrant the increased costs. It is true that apparatus tests have considerable face validity and thus are meaningful to applicants. Nevertheless, paper and pencil tests of the sort used in the present investigation, while novel to the large bulk of applicants, were readily accepted by them. By means of proper instructions and judicious choice of items paper and pencil tests, too, can be made with satisfactory face validity.

The battery consisting of the dotting, tapping, judgment of distance, distance discrimination, and mechanical principles tests seems most promising as a selective device for the selection of safe vehicle operators. On two different groups of operators it has given acceptable results. One difficulty with the battery is that most weight is assigned to two very short tests. It is almost inconceivable that the dotting and tapping tests with their half-minute times could have reliability coefficients higher than .60, and, indeed, a lower figure would seem more reasonable. It is possible, of course, that they do measure abilities of particular importance to safe operation, but certainly more evidence than two studies covering only a total of some 220 cases is necessary before any great confidence can be placed in them.

Another difficulty with the battery utilized here, and in fact, with any paper and pencil tests, is their relative uselessness for training purposes. Scores on visual and reaction time tests are most helpful aids in safety training programs. Having estimates of an individual's visual acuity glare sensitivity, reaction time, etc., intelligent recommendations can be made and pertinent training initiated to compensate for any deficiencies. With motor vehicle operators Fletcher has been able to produce significant and long term improvements in driving safety by utilizing such tests and interpreting the significance of the scores to the individuals concerned (3). Probably the best that paper and pencil tests can do for training is to indicate which individuals should be given special attention in any safety training programs.

Summary

The scores earned by 67 taxicab drivers on eight paper and pencil tests and an interest inventory, together with certain personal data items, were studied in relation to safety of operation. Accidents during the first five weeks of employment formed the criterion in the computation of validity coefficients. Dotting and tapping tests were found to have the highest validity. Tests involving judgment of distances and of knowledge of simple mechanical principles yielded low validity coefficients. A combination of the scores on the five foregoing tests, weighted on the basis of evidence collected in another investigation with an entirely different group, yielded a battery which was found to have a validity coefficient of .59 for accidents. An arithmetic test and a paper and pencil test of complex reactions were found to be useless in predicting accidents. Interest measures showed some promise, particularly for scales of occupational level, outside occupations and related occupations. No significant relationships were found between the accident criterion and age, education, and previous driving experience.

References

1. Brown, C. W., and Ghiselli, E. E. Accident proneness among streetcar motormen and motor coach operators. *J. appl. Psychol.*, 1948, **32**, 20–23.
2. DeSilva, H. R. *Why we have automobile accidents.* Wiley, 1942.
3. Fletcher, E. D. Capacity of special tests to measure driving ability. Mimeographed, undated.
4. Ghiselli, E. E., Brown, C. W., and Minium, E. W. The use of test scores for the prediction of accidents of street car motormen. Report to the Municipal Railway System of San Francisco, 1946.
5. Ghiselli, E. E., and Brown, C. W. *Personnel and industrial psychology.* New York: McGraw-Hill, 1948.
6. Personnel Research Section, Adjutant General's Office. Statistical Manual. 1944.
7. Viteles, M. S. *Industrial psychology.* New York: Norton, 1932.

Some Precautions in the Use of the Per Cent Method of Job Evaluation

William D. Turner

University of Pennsylvania

Section 5 of the computation for establishing factor comparison scales by the per cent method of job evaluation (3) involves a table of $F\%/J\%$ ratios (or estimates of relative job totals, see 4, p. 156F) the columns of which have different totals. It may be assumed that such differences between column totals follow principally from fortuitous variation of intra- and inter-job factor patterns, since judgment errors contributing to such differences will tend to be cancelled out in each column total. The individual ratios in these columns are quite fallible estimates of relative job totals, because each ratio is based ultimately on two independent and fallible per cent ratings. An improved estimate of job totals could be had if there existed several such estimates for each given job, which could then be averaged with the expectation of cancelling out material portions of the judgment errors inhering in each. One's first impulse is to average the ratios in each given *row* of Section 5. But the aforementioned differences between the totals of the *columns* of these ratios indicate that each column is of a different order of magnitude, and that an estimate in one column is therefore not strictly comparable with one in the same row but another column. Accordingly, in Section 5 of the computation, four of the columns are multiplied by "Reduction Constants" so that their respective totals become equal to that of the remaining (M) column whose total is arbitrarily taken as a base. The results of such multiplications appear in Section 8 of the computation. The *rows* in Section 8 are then summed (which amounts to averaging, since each row contains the same number of ratios) to yield the column of "(8) Totals" appended to Section 4. These "(8) Totals" represent the improved estimates of job totals sought above.

Likewise, improved estimates of factor totals are obtained by applying "Reduction Constants" to the *rows* of Section 6, $(J\%/F\%$ ratios), so that the resulting *columns* in Section 7 contain relative factor totals of comparable orders of magnitude, which are in turn summed to obtain improved estimates of such totals. See the "Total" row in Section 7, whose contents are subsequently "converted" to adjust their general level of magnitude to that of the "(8) Totals" discussed above, before being appended to Section 3.

Consequences of Eliminating the "Reduction" Process

Hay (2) essentially proposes to eliminate the "Reduction" operation in question, and, hence, to average incomparable relative figures. If one had valid reason to weight differently, say, the columns in Section 5, one could quite properly depart from the writer's procedure. But there is no known rational basis for such differential weighting, and Hay's procedure accepts an irrational and adventitious weighting for which there is no theoretical defense.

Whether Hay's method can be justified in practice depends then on the magnitude by which his results depart from those yielded by the writer's method, and on the relative value of committee time lost in correcting the larger, ultimately detectable errors introduced by Hay's method.

When Hay's abbreviated procedure is applied to the data in Section 5 of the writer's article (3), and when the general level of magnitude of the resulting relative job and factor totals is adjusted to that of such totals in Sections 3 and 4 of the method, factor ratings by $F\%$ and $J\%$ comparable with those in Section 11 may be computed. When this is done, and when such totals and ratings are compared with those obtained by the complete computation, 45% of the 137 percentage differences between the results by the two methods are greater than zero. 28% of these percentage differences exceed 3; 21% exceed 4; 20% exceed 5; and 13% exceed 6. Three each of these differences exceed 7 and 8; and one each exceeds 9, 10, 11, and 12.

The significance of the foregoing percentage differences depends particularly on the frequency and magnitude of the larger ones relative to a job rating committee's judgment error. So far as the writer is aware, no one has determined the exact value of such an error for a given committee. However, the committee whose results illustrate the writer's article showed an average 5% percentage difference between its original and its reviewed ratings on about 300 jobs, when the review in question followed soon after a leveling off of the committee's skill. Percentage differences between the results of a subsequent review and the one mentioned can safely be estimated to lie between 2% and 4%. Assuming a 3% error as a compromise figure, about a third of the ratings which Hay's procedure would have supplied to this committee would have departed from the ratings produced by the original computation by amounts equalling or exceeding the committee's own emerging average judgment error, and several of them would have differed from the latter ratings by as much as three or four times this error.

Again, the foregoing discussion obviously assumes that ratings produced by the original computation approximate true rating values more

closely than do those produced by Hay's computation, and that the size and frequency of discrepancies between the ratings produced by the two methods signify the degree of fault in ratings produced by Hay's method. This assumption is made because, as noted above, Hay's computation involves the operation of averaging relative figures of different orders of magnitude. Such an operation augments the inevitable errors (of judgment) already present in the ratings produced by the original computational procedure. The rational "Reduction" process in Sections 5 and 6 of the original procedure minimizes this difficulty.

Statisticians will question why an *average* percentage discrepancy of 1.8% between the results by the two computational methods is important when the committee's *average* judgment error itself is probably as great as 3%. Since a job evaluation committee is not so concerned with achieving a small *average* error characteristic of its ratings in the aggregate as it is with minimizing its error in each particular rating it makes, those large errors actually introduced by Hay's method, and which the committee can come ultimately to detect, become important. There are 18 such discrepancies in the illustrative case, which are at least twice as great as the committee's average judgment error, and 4 of which are at least three times as great as this error. Only a few such errors in the original scales are all that are needed to throw many subsequent ratings out of line before the committee can become able to recognize their incorrectness, and the committee would then need considerable time to make the necessary corrections. Such committee time would be considerably more expensive than the hour or less gained by the computer using Hay's abbreviated procedure; and unless the discrepancies in question are ironed out, errors and borderline cases in job grading will be increased.

On the basis of the foregoing findings and conclusions, the writer sees fit to warn against the use of Hay's abbreviated procedure, and to recommend that the complete computation as originally set forth be followed in all cases.

Mathematical Relations Underlying the Per Cent Method

It should be observed that Hay's proofs that row totals correspond with job totals, and that reciprocals of column totals correspond with factor totals, obscure their underlying assumption that relative values of differing orders of magnitude can be validly averaged. The writer's present findings, and the marked differences between some of the "Reduction Constants" which appear in Sections 5 and 6 of the writer's article (3), emphasize the practical untenability of such an assumption and of the proofs which Hay bases on it.

In the writer's second article (4) the attempt was made to show some of the meaning of the per cent method's complete computations, without recourse to algebraic notation. An algebraic version of pages 155 and 156 of this article follows:

Let

r = any one of a number of factor ratings obtained during the use of established factor rating scales;

T_J = the total of factor ratings (r's) for any job;

T_F = the total of factor ratings (r's) for any one factor for any given group of jobs for which T_J's are also available;

$J = r/T_J$ = a "J value" (see Table II, p. 155, Reference 4) corresponding to a given r; and

$F = r/T_F$ = an "F value" (see Table III, p. 155, Reference 4) corresponding to the same r.

Then, $F/J = (r/T_F)/(r/T_J) = (r/T_F)(T_J/r) = T_J/T_F$.

But for any given factor grouping of r's, T_F would be constant, and may be regarded as equal to unity. Therefore, the *relative job totals*, or $T'_J = F/J$.

By corresponding proof, the *relative factor totals*, or $T'_F = J/F$.

An algebraic account of the more complex situation that arises when fallible $F\%$ and $J\%$ values are substituted for mathematically infallible F and J values, would be essentially descriptive. Since the algebraic expressions involved would be even less easily followed by most readers than is the verbal account in the writer's second article, the writer has not published such an algebraic description.

Weber's Law and Job Evaluation

Hay implies (1, 2) that Weber's Law applies to factor comparison job evaluation. In brief, Weber's Law says that a discriminable difference between two physical stimuli bears a constant ratio to the level of magnitude of the stimuli themselves. The absence of any physical or objective measure of job values makes an application of Weber's Law to job evaluation data logically impossible. Hay refers to unpublished evidence of his own, to the effect that discriminable differences expressed in rating scale units bear constant ratios to the magnitude of ratings in factor comparison job evaluation, but apparently fails to consider the subjective nature of the scale units in which such differences *and* their corresponding ratings are necessarily expressed. Hence, the law which can properly express Hay's observations is not Weber's psychophysical Law, but a psychological Law of Per Cent Judgment which may be

written, $L/J = K$, in which

L = any discriminable difference (limen) expressed in per cent rating scale units or their equivalent;[1]

J = the point level of the scale rating characterized by L; and

K = an empirical constant.

It is evident that this Law of Per Cent Judgment involves no physical or objective measures, and that any similarity which it bears to Weber's Law is essentially mathematical; it is a psychological and not a psychophysical law. This latter distinction, which may seem to be purely theoretical, has an important practical implication for job evaluation. An assumption that Weber's Law applies to job evaluation data implies quite incorrectly that job values can be ascertained by some physical or objective method of measurement. Such an assumption would lend unwonted support to one of job evaluation's recurrent delusions which holds that "somewhere there exists the 'real objective truth' about job values." The Law of Per Cent Judgment emphasizes the inescapably subjective nature of the process of job evaluation.

Hay's proposed "limen" of 15% is actually determined by the Law of Per Cent Judgment formula given above, with L equalling the scalar distance from the *committee's* corresponding final ratings which includes at least 75% of committee members' individual preliminary ratings; with J equalling the former ratings; and with the K value multiplied by 100 to yield a percentage figure. However, the members of one of the rating committees whose work the writer has directed manifested in the aggregate a corresponding "limen" ratio equal to 9%, with corresponding "limens" for individual members lying between 6% and 11%. In order that a committee's judgments may express more fully the accuracy of which the committee and its members are capable, the writer recommends again (cf. 3) that the size of geometric steps on factor rating scales be commensurate with a committee's emerging judgment accuracy rather than with a prescribed standard "limen" of 15%.

Received March 31, 1949.

[1] The term, "their equivalent" is meant to signify factor rating scales derived by the per cent method, or such scales as derived by Benge's method. The latter equivalent is justified by the extremely close correspondence (apparently limited only by judgment error) between ratings by the per cent method and Benge's method. Calling the present law one of per cent judgment relates it directly to the more controllable process of (per cent) judgment rather than to the salaries of equitably paid key jobs from which Benge derives his scales. The generally linear correlation between either per cent method or Benge method ratings and non-negotiated rates of pay, and the much closer agreement between ratings by the two methods in question, indicate that pay rate setting in the absence of job evaluation rests on a less accurate form of per cent judgment.

References

1. Hay, Edward N.   Characteristics of factor comparison job evaluation.   *Personnel*, 1946, **22**, 370–375.
2. Hay, Edward N.   Creating factor comparison key scales by the per cent method. *J. appl. Psychol.*, 1948, **32**, 456–464.
3. Turner, William D.   The per cent method of job evaluation. *Personnel*, 1948, **24**, 476–492.
4. Turner, William D.   The mathematical basis of the per cent method of job evaluation.   *Personnel*, 1948, **25**, 154–160.

Predicting Subject Grades of Liberal Arts Freshmen with the Kuder Preference Record *

Dorothy Terry Hake and C. H. Ruedisili

University of Wisconsin

The solution to the problem of the prediction of achievement is still in a preliminary stage. Interests are generally conceded to be important in determining success in any field of endeavor, but they are only one of many factors, such as intelligence, attitudes, and personality traits. The present study deals with the relationships between college achievement in specific subject areas and interest test scores. We have investigated the value of the Kuder Preference Record in predicting the first-semester grades made by freshmen at the University of Wisconsin.

Since the Preference Record was developed to measure interests, a high correlation with specific achievements would not necessarily be expected, but a low positive correlation with general college achievement has been found in previous studies. Crosby (2) has found positive correlations in the high .60's between the Preference Record Scientific scale and chemistry and biology grades, and between the Computational scale and accounting grades. His subjects were students scoring above the 90th or below the 10th percentile in each of the Preference Record scales, and therefore the relationship between interests and grades is not as high as it appears at first glance. If the total distribution of interest scores had been used, the correlations would have been considerably lower (6). Thompson (9) reports some success in predicting dental school success by using the Preference Record in conjunction with the MacQuarrie Test for Mechanical Ability. Bolanovich and Goodman (1) used the Preference Record to select women students for training programs in electronic engineering during the war. They found significant differences between successful and unsuccessful students. The former showed high scores on the Computational and Scientific scales, and low scores on the Musical and Clerical scales. Yum (10) found differences among students enrolled in the Physical, Biological, and Social Sciences, and also found low positive correlations with grade-point average.

Strong (8) mentions that since college courses are largely elective, the student does not choose courses in which he has no interest. He points

out that interests, then, would not be important in determining success in elective college courses. University of Wisconsin freshmen at the time of this study, however, were much restricted in their choice of courses. The subject fields used were all freshman courses, and the usual freshman program consisted of four out of the five areas: English, Foreign Language, History, Science, and Mathematics. The problem of free-selection, then, is largely eliminated from the study and, presumably, measured interest scores might have some part in determining success in these subjects.

Procedure

The Preference Record was given to all Letters and Science freshmen who entered the University of Wisconsin in the fifteen-weeks summer session and the fall semester of 1943. The first-semester grades in five subjects and the over-all grade-point averages were obtained. The subjects included were English (N = 579), Science (N = 528), History (N = 402), Foreign Language (N = 477), and Mathematics (N = 201). All freshmen who took the Kuder Preference Record in the entrance examinations and completed the first semester, including one or more of the five courses, were included in the study; altogether, 594 students met these requirements (3). Men and women students were combined in this group since there were no separate norms for the Kuder Preference Record at the time this study was made (4).

Results and Discussion

The means and standard deviations of the Preference Record scales and of grades were computed for each subject area and for the whole course load (Grade-Point Average). These figures are in terms of raw scores for the scales and are not equated from scale to scale. The grades are in terms of grade-point averages, with 1.00 representing a grade of C. Differences in mean interest scores between subject groups are frequently large. The Mathematics group seems to be especially deviant in interest scores. It is relatively high in Computational and Scientific interest (39.32 and 66.89 respectively) and low in Social Service (69.17). This suggests that students taking Mathematics may have had more specific interest in the subject than did students taking other subjects. This seems likely, since Mathematics is often avoided as a difficult subject. Interest scores for the other subject groups are fairly similar, however, bearing out the preliminary hypothesis that these students' interests do not materially affect their choice of courses. The highest grade-point average was obtained by the Foreign Language group (1.37) and the lowest (1.15) by the English students. The highest S.D. occurred in the History group (1.46) and the lowest (0.82) in the English group.

Table 1
Correlations between Preference Record Scores and Grades

Scale	English	Science	History	Foreign Language	Mathematics	Grade-Point Average
Mechanical	−.15	.04	−.13	−.18	−.02	−.07
Computational	−.10	.10	.03	.03	.10	.04
Scientific	−.04	.18	.04	−.02	.10	.07
Persuasive	.08	−.05	.11	.00	−.02	.09
Artistic	.02	−.03	−.03	.01	−.02	.11
Literary	.25	−.01	.13	.12	.02	.13
Musical	.03	−.06	−.01	.10	−.10	−.01
Social Service	−.03	−.14	−.14	−.09	−.06	−.12
Clerical	−.07	−.04	.04	.07	−.01	.00

Table 1 contains the correlations between Preference Record scores and subject grades as well as grade-point average. The first column consists of the correlations between the grades of students who took English and their scores on each of the nine scales. The other columns show the same relationships for the remaining subjects. There are some expected relationships apparent in this table. The highest positive correlations are between Literary interests and grades in English, History, and the general grade-point average. The correlation between Science and the Scientific scale is in the expected direction. The inverse correlations between Mechanical interests and grades in English, History, and Foreign Language also seem reasonable. The Artistic and Clerical scales have the lowest correlations with grades in general.

The correlations among the individual scales on the Preference Record for the group as a whole are shown in Table 2. These intercorrelations, on the whole, are either low positive or negative. Relatively high positive correlations, however, are found between the Scientific and Mechani-

Table 2
Intercorrelations among Kuder Scores

	Mechanical	Computational	Scientific	Persuasive	Artistic	Literary	Musical	Social Service
Computational	.25							
Scientific	.64	.34						
Persuasive	−.20	−.19	−.35					
Artistic	.04	−.28	−.12	−.11				
Literary	−.31	−.15	−.31	.28	−.08			
Musical	−.28	−.18	−.23	.07	.16	.10		
Social Service	−.27	−.24	−.19	.09	−.10	−.02	−.02	
Clerical	−.20	.40	−.29	.07	−.20	.04	−.04	−.08

cal scales, the Literary and Persuasive scales, and between the Clerical and Computational scales. The highest negative correlation is that between the Scientific and Persuasive scales, while other fairly large inverse correlations are those between Mechanical and Literary, Scientific and Literary, Musical and Mechanical, Computational and Artistic, Scientific and Clerical. These same tendencies also can be noted in the intracorrelations found in five other groups, as described in the *Revised Kuder Preference Record Manual* (5).

The results of the Wherry-Doolittle Method of Test Selection included the progressive shrunken multiple-correlation coefficients, the uncorrected multiple-correlations, the name of the first scale which was not included in the battery because it caused a decrease in the shrunken multiple-correlations, and K (the coefficient of alienation). The highest of these shrunken multiple-correlations is only .3093 for General Grade-Point Average. One might expect to find that general school achievement can be predicted more accurately from interests than can grades in any one subject. Since the multiple-correlations for English (.2936) and History (.3022) are almost identical with that for the Grade-Point Average, however, there must be other factors involved. Possibly the relatively large size and heterogeneity of these groups may account for the higher multiple-correlations. The lowest multiple-correlation (.1997) is that using Mathematics as a criterion, and this group has the smallest N. It must also be remembered that this group seems to have relatively similar interests, and thus is a more homogeneous group. For the whole group, the test adding more chance error than validity was the Clerical scale; for the History, English, and Foreign Language groups it was Artistic; for Science it was the Persuasive scale, and for Mathematics the Computational scale.

Discussion

In general, it seems likely that interests, as measured by the Kuder Preference Record, are a relatively minor factor in predicting college achievement. Used alone, the Preference Record would probably be of little help. The addition of the Preference Record to college entrance examination batteries may be advisable, however, since interest measures may very well contribute significantly to the multiple-correlation obtained with the traditional aptitude and achievement tests. Further research which combines the Kuder Preference Record and other interest measures with aptitude and achievement scores is advisable before interest tests are rejected as being useless in predicting college achievement.

Summary

Scores on the Kuder Preference Record were compared with the grades obtained in five subject-fields by 594 students. The students

were first-semester freshmen in the College of Letters and Science at the University of Wisconsin. The correlations between grades and grade-point averages and scores on the Preference Record scales were computed and from these the shrunken multiple-correlations were obtained for each subject group and for the grade-point averages of the whole group by means of the Wherry-Doolittle Method of Test Selection.

1. The means and standard deviations of the raw Preference Record scores, with one exception, did not differ widely with respect to subject groups, indicating that interests are not an important factor in choosing freshman courses, and thus presumably are important in determining success.

2. The correlations between the scales and the grades and grade-point average are, on the whole, fairly low, but certain logical relationships between scores on the Preference Record and the subject groups can be noted. The Literary scale was found to have the highest positive correlations with the subject grades and grade-point average, while the Mechanical and Social Service scales showed fairly high inverse correlations.

3. The intercorrelations among the scales were low and negative, on the whole, but fairly high positive intercorrelations were found among the Mechanical, Computational, and Scientific scales, between the Literary and Persuasive, and between the Clerical and Computational scales. These same tendencies have been noted in other studies.

4. The results of the Wherry-Doolittle method show that a few of the scales, such as Mechanical, Scientific, Literary, and Social Service, were more useful than others in contributing to the multiple-correlations. The resulting shrunken multiple-correlations were all low. The largest obtained was that for the total grade-point average (.3093), and this approximated the best subject-fields, History (.3022), and English (.2936). It is concluded that interests, as measured by the Kuder Preference Record, may play a minor role in determining school achievement. In conjunction with other tests (achievement, scholastic aptitude, intelligence, attitude, etc.) the scores of this test may prove useful in the prediction of college grades.

References

1. Bolanovich, D. J., and Goodman, C. H.   A study of the Kuder Preference Record.   *Educ. psychol. Measmt.*, 1944, 4: 315–326.
2. Crosby, R. C.   Scholastic achievement and measured interests.   *J. applied Psychol.*, 1943, 27: 101–104.
3. Froehlich, G. J.   The prediction of Academic Success at the University of Wisconsin, 1909–1941.   *Bulletin of the University of Wisconsin. Bureau of Guidance and Records of the University of Wisconsin, October* 1941.

4. *Intermediate manual for the Kuder Preference Record,* Chicago: Science Research Associates, 1944.

5. Kuder, G. F.   *The revised manual for the Kuder Preference Record,* Chicago: Science Research Associates, 1946.

6. Peters, C. C., and Van Voorhis, W. R.   *Statistical procedures and their mathematical bases.*   New York: *McGraw-Hill Book Co.,* Inc., 1940.

7. Stead, W. H., and Shartle, C. L.   *Occupational counseling techniques.*   New York: American Book Co., 1940.

8. Strong, E. K.   Vocational interests of men and women.   *Stanford University Press,* 1943.

9. Thompson, C. E.   Personality and interest factors in dental school success.   *Educ. psychol. Measmt.,* 1944, 4: 299–306.

10. Yum, K. S.   Student preference in divisional studies and their preferential activities.   *J. Psychol.* 1942, 13: 193–200.

MMPI Personality Patterns for Various Occupations

E. E. Daniels and W. A. Hunter

VA Regional Office, Phoenix, Arizona

Two closely connected provocative questions regarding the role of dynamic personality patterns[1] in relation to vocational selection and personnel placement prompted the present investigation.

First, are there rather definite personality patterns which tend to gravitate toward certain of the multitude of occupations in the vocational world? Second, are there rather fixed "personality demands" in the various occupations which make up the work of man?

Industrial and employment interviewers, guidance personnel, teachers, etc., are aware of the numerous expressed reasons why individuals have gone to work in a particular job or embarked on a certain career. Menninger points up the problem as follows: "It would be interesting to examine how it came about that some people must do continuously what seems chiefly drudgery, while other people are able to do what seems to be pleasurable and even delightful work, if indeed it can be called work at all" (1). Is there a particular optimal pattern of personality factors for each occupation which when met contributes to occupational success and satisfaction? Will a standardized personality test be subtle enough to ferret out these patterns for research studies and even for practical application in job counseling?

A review of literature revealed a study made by Harmon and Wiener (2) who asserted, in using the MMPI, that personality characteristics appear to be of crucial importance in the actual choice of a vocation, a contention which appears to be a distinct aid in the prognosis of the success in training. In another study Verniaud (3) administered the MMPI to clerical workers, department store saleswomen and optical factory workers and concluded that saleswomen tend to make responses designated as "masculine," industrial women show definite trends toward hypomania and psychasthenia, while the clerical workers approach more nearly those responses which had been termed "normal." The present study used the MMPI (4) as an exploratory tool in an attempt to determine whether a relationship exists between the total personality "work needs" and the "personality demands" of occupations.

[1] The term "pattern" is used in the dynamic sense and substituted for the term "deviate" in the MMPI.

Methodology

The dynamics of this study were worked on over a period of 32 months in the Veterans Administration Regional Office, Phoenix, Arizona, and the raw data were collected from four VA Guidance Centers over the State. The conclusions are based on material drawn from a study of 893 veterans under both Public Laws 16 and 346. All cases had availed themselves of complete advisement and guidance, as set forth in the VA Manual of Advisement and Guidance (5), which culminated in the veteran's choice of an occupation. The MMPI categories were coded along with veteran's name on an IBM card, and the occupations which covered 97 category groupings from the Dictionary of Occupational Titles (6) were then obtained by IBM selection in terms of DOT code number from the master file of status cards. All cumulations were made on the IBM.

The group represented males of an average age of 23 years, constituting several racial groups from all sections of the United States and 90 per cent were high school graduates. For the purpose of coding on the IBM cards, the range of T scores on the MMPI was grouped at the center of a 10-point spread, i.e., at 55 for the range of scores 50 to 60.

After averages had been determined for each occupation and for each scale on the MMPI, tables were made up for each personality dimension separately. An F-test was made for the various MMPI characteristics. The F scores for the Mf, Pd, Sc, and Ma respectively were 5.85, 5.24, 2.79, and 3.36, which was very significant for Mf, Pd, and Ma, and significant for Sc, which means that the chances are less than 1 to 99 that so large an F could have occurred in a really homogeneous population. All semi-skilled and unskilled occupations, DOT codes 6-00 through 9-99, were omitted, leaving a total of 67 occupations. Using Fisher's Small Sample statistical technique, the significance of the obtained difference in the means between various pairs of occupations was calculated. In those cases where the difference was found significant there are about five chances, and for very significant about 1 chance in 100 that this could have occurred by random sampling. Not all pairs where there is a significant or very significant difference between the means are represented in the tables; only four of the MMPI scales out of a total of nine scales were selected for presentation in this article.

Results

Table I is composed of a list of selected occupations in terms of the average T score on the Masculine-Feminine, Psychopathic, Schizophrenic

Table 1

Mean T-scores on Four Scales of MMPI for Selected Occupational Groups

Occupational Group	N	Mf Scale	Pd Scale	Sc Scale	Ma Scale
Social Scientist	5	75.0	63.0	51.0	69.0
Physician	5	69.0	65.0	55.0	69.0
Author, Editor, Reporter	10	67.0	70.0	56.0	67.0
Chemist	7	65.0	56.0	52.0	57.0
Draftsman	13	58.8	57.0	55.8	57.0
Farmer, Livestock	5	47.0	57.0	55.0	55.0
Accountabt	13	61.0	55.0	58.0	50.0
Medical Technician	8	55.0	63.0	48.0	56.0
Insurance Salesman	10	57.0	63.0	51.0	62.0
Athletic Coach	10	53.0	66.0	44.0	55.0
Auto Mechanic	33	52.2	62.0	52.0	61.0
Barber	4	65.0	55.0	52.0	57.0
Dancing Instructor	10	62.0	59.0	56.0	62.0
Farmer General	15	49.0	58.0	52.0	56.0
Commercial Artist	9	51.0	65.0	47.0	58.0
Social Worker	4	60.0	52.5	47.0	55.0
Clerk General	14	53.0	54.2	40.0	56.0
Auto Body Repairman	19	55.0	52.9	49.0	60.0
Managers	6	52.0	50.0	52.0	55.0
Lawyer	13	64.0	61.0	54.2	64.2
Typist	3	58.0	62.0	45.0	58.0
Radio Announcer	4	60.0	62.0	57.0	67.5
Electrical Repairman	7	52.0	58.0	52.6	49.3
Auto Upholsterer	6	57.0	63.0	55.0	53.3
Electrician	10	53.0	56.0	52.0	53.0
Kindergarten Teacher	5	57.0	63.0	55.0	63.0

and Manic scales. Tables 2, 3, 4, 5, and 6 giving complete results are omitted in this article because of cost.[2]

Attempts were made originally to group the accupations according to the major occupational groupings in the Dictionary of Occupational Titles. These broad groupings due to lack of homogeneity obscured the personality patterns of occupations within the group while showing no differences between groups. While occupational groupings may be made on the basis of one scale, it tends to obscure differences on other scales.

The data indicate that the means on the MMPI scales for the various occupations tend to scatter rather widely about the T score of 50; whereas, if the mean of all occupations combined is calculated, the mean approaches

[2] For Tables 2, 3, 4, 5, and 6 order Document 2694 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting $0.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or $1.00 for photocopies (6 × 8 inches) readable without optical aid.

rather closely the T score of 50 for each scale.   The data would appear to
indicate significant differences between the means of personality patterns
as related to the various occupational objectives.   Perhaps this would
indicate that extensive differences in personality patterns exist between
occupational groups at the extremes of the distribution of occupations
obtained on the MMPI scales.   It is found, for example, that the mean
of personality scores for occupations taken from near the middle of the
distribution is very significantly different from the means at either ex-
treme.   It is also noted that, since the MMPI scores for the various
occupational groups tend to spread as indicated in Table 1, the statisti-
cally not significant differences indicate a degree of difference that possibly
may be considered as placing each occupation near its optimal level on
the scale for this particular personality pattern of the MMPI.

In contrast to a rather common interpretation of the MMPI that a
score below 70 does not indicate a significant personality deviation, it is
believed that any individual deviation from the mean T score, either posi-
tive or negative, on any personality scale is indicative of a certain tendency
toward behavior in that direction, and that extremes, such as a critical T
score of 70, are not necessary for the instrument to have definite meaning
and application in the industrial field.

The *Masculine-Feminine* pattern on the Minnesota Multiphasic would
appear, from these findings, to indicate a "work need" often requiring
rechanneling in order that the occupational satisfaction of the basic
Masculine-Feminine content of the total personality be achieved.   This
may be illustrated, for example, in the statistical difference between social
scientist and farmer, livestock.   The "work need" for a social scientist is
an understanding of the problems of other human beings, the prerequisite
of which is a high degree of sensitivity as seen in the Masculine-Feminine
pattern.   Another example of this need for redirection of the Masculine-
Feminine component of the total personality can be seen in the occupation
of physician, whose "work need" is expressed in his "bedside" manner, as
compared to the low degree of this pattern in the occupation of draftsman,
the difference between which is significant statistically.   This dependency
of occupational choice upon the Masculine-Feminine level as indicated
by the MMPI has been empirically tested numerous times by us during
vocational advisement and guidance by attempting to get a person who
was interested in some occupation such as barber or beautician to consider
the objectives of meat cutter or butcher, or vice versa.   In all cases there
has been a violent rejection of the consideration of the alternate objective.
Statistical evidence and protocols seem to indicate that professions of a
so-called highly cultural nature require as a fundamental "work need" a
degree of Masculine-Feminine pattern approaching 70 T score on the
Minnesota Multiphasic.

The *Psychopathic pattern* of the Minnesota Multiphasic would appear to be characterized by aggressiveness or by asocial behavior. Underneath this aggressiveness is the raw hostility and destructiveness as demonstrated by ample clinical evidence. This hostility originates from childhood reactions to authority in the family situation. Menninger points out that "The concept of work as drudgery which everyone experiences to some extent and which some persons experience to a very high degree, is bound up with this resistance to authority" (1).

The "work needs" of the individual personality with a high degree of Pd pattern may be illustrated in the choice of occupation such as author, editor, reporter, or athletic coach. "Purposeless destructiveness and aggressiveness may be molded and guided into the constructive activity of work" (1). This hostility may also be observed in those cases characterized by failure due to the unrecognized "work need" in the high Pd pattern of personality, as, for example, in the case of the veteran who was striving to achieve a father identification in his occupational selection by entering the same field as his father. The veteran was unsuccessful in his efforts until returning to the psychologist, wherein it became apparent from the evidence that hostility between the veteran and his father dated back to the earliest years. In attempting to alleviate the occupational blocking a counseling technique was applied whereby the veteran was urged to return East and spend his entire time, if possible, in close companionship with his father. This recommendation was accepted, and the veteran a number of weeks later returned with renewed interest and determination to succeed in his occupational efforts. The removal of the cause of competition with the father resulted in a freer expression of his occupational efforts in the same field. This would appear to illustrate the blockage of "work needs" which must be recognized if a rechanneling of childhood hostility into an occupational goal is to be successful.

Statistical findings in these data indicate that the difference between the occupation of athletic coach and the occupation of manager is significant, which might be interpreted to mean that the managerial occupations require a complete rechanneling of hostility in the direction of objectivity in management and administration on a more socially accepted level, whereas the athletic coach utilizes his aggressiveness mostly on a work level of physical effort (sports).

The *Schizophrenic pattern* on the Minnesota Multiphasic in this material appears to indicate a "work need" wherein the individual does not have to associate too closely with other people. The mechanism of isolation is well delineated in the clinical syndrome of Schizophrenia. The same mechanism appears to be effective to a lesser degree in influencing persons with a high Sc score in the choice of occupations, as exemplified in the difference between the draftsman as compared to athletic

coach, statistically a very significant difference. The occupation of draftsman may be considered as an isolating occupation, whereas the occupation of athletic coach required an essential capacity for dealing closely with other people. Another example which may be cited shows the difference in "work needs" for the occupation of typist, which is essentially a social and interpersonal occupation, as compared with the draftsman. Thus, occupations indicating a significant low degree of Sc pattern would appear to require a "work need" wherein the individual may satisfy his gregariousness, as compared to a high degree of Sc pattern wherein the occupation requires little association with others in the work situation.

The "work needs" as indicated in the *Manic pattern* of the Minnesota Multiphasic would appear essentially to be an outlet for enthusiasm and a high degree of overt activity. In the occupation of radio announcer, as compared to electrician, the difference is significant statistically, and may afford statisfaction in the occupation by providing an outlet for emotional and verbal expression. Again, this "work need" is clearly illustrated in the occupation of teacher, kindergarten, as compared with the occupation of electrical repairman, where the rechanneling of emotional content may be observed. Occupations which require dynamic behavior, such as that of lawyer, is another example, as contrasted with automobile uphoslterer, the difference between which is statistically significant.

Thus, many occupations seem to utilize and perhaps demand a personality pattern in which there is a great deal of spontaneity and enthusiasm expressed, whereas other occupations make little use of this personality pattern. Lewis' study on this problem asserted "that there is a relationship between occupational interests and personality tendencies" (8).

Discussion and Indicated Application

The "work needs" of the total personality have been presented in this investigation with the intent of stimulating further research. The dynamic relationship between the "work needs" of the total personality and the selection of occupation would appear to us to be significant. From these findings it seems desirable to scrutinize closely occupations in terms of "personality demands." *These "personality demands" once established could then be matched with the "work needs" of the total personality as indicated on the Multiphasic patterns in a manner similar to that already established in the occupational realm of job demands and physical capacities analysis.*

In this study the dynamics of the total personality are viewed in terms of their psychogenetic origins and their development through conditioned response, a learning phenomenon which may be easily observed in the mechanism of parental identification.

By utilizing this technique and viewpoint it seems to the authors that the Minnesota Multiphasic is a fairly sensitive instrument for measuring the total personality "work needs" in relation to the suitability of occupations having certain "personality demands."

References

1. Menninger, K. *Love against hate.* New York: Harcourt, Brace and Company, 1942, pp. 136–137.
2. Harmon, L. R., and Wiener, D. N. Use of the Minnesota Multiphasic Personality Inventory in vocational advisement. *J. appl. Psychol.,* 1945, 29, 132–141.
3. Verniaud, W. M. Occupational differences in the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.,* 1946, **30**, 604–613.
4. Hathaway, S. R., and McKinley, J. C. *Manual for the Multiphasic Personality Inventory,* New York: The Psychological Corporation, 1943.
5. Scott, I. D. *Manual of advisement and guidance.* Washington, D. C.: Veterans Administration, U. S. Government Printing Office, 1945, pp. 11–52, 83–180.
6. *Dictionary of occupational titles,* **U. S.** Department of Labor and U. S. Employment Service, Part I, Definitions of titles; Part II, Titles and codes; Part IV, Entry occupational classification, and supplement edition III, Washington, D.C.: U. S. Government Printing Office, 1939.
7. Lewis, J. A. Kuder preference record and MMPI scores for two occupational groups. *J. consult. Psychol.,* 1947, 11, 194–201.

Correcting Special Ability Test Scores
for General Ability

Abraham S. Levine

University of Minnesota

Most paper and pencil tests designed to measure special abilities or aptitudes correlate positively in varying degrees with tests of general ability or intelligence. This fact does not particularly detract and may actually enhance the predictive efficiency of special ability tests for occupations in which success is positively related to general ability. However, there are a large number of jobs particularly in the semi-skilled trades which do not require more than a modest level of general intelligence and for which a high degree of such ability has actually been shown to be related to high turnover rates. In these occupations the best predictors for the most part have been apparatus tests which are negligibly correlated with tests of general or verbal intelligence. For reasons of economy it is desirable wherever possible to administer group paper and pencil tests rather than individual apparatus tests. Therefore, if the effect of general ability could be partialed out, the utility of these contaminated paper and pencil tests as guidance and selection instruments for the relatively low I.Q. occupations may be increased.

The proposed method of correcting for the effect of general intelligence in a special ability test represents a simple application of the regression coefficient. A regression coefficient enables one to estimate the scores on a test if one knows the scores on another test with which it is correlated and the magnitude of this correlation coefficient for a given sample. For the sake of illustration, let us choose two tests: (1) a general ability test such as the Tiffin and Lawshe Adaptability Test; and (2) a special ability test such as the Bennett Test of Mechanical Comprehension. Let us say that John Black obtained a score on the Adaptability test which was one and one-half standard deviations above the mean of a specified group, and there was a $+.50$ product moment r between the Adaptability and Mechanical Comprehension tests for this group. The best estimate of John's score on the Mechanical Comprehension test would be $.50 \times (+1.5)$ or a standard score of $+.75$.

The proposed correction for general ability substracts or adds to the special ability test score an amount defined by the size of the regression coefficient and the deviation of a general ability test score from the mean.

Thus, for example, if John Black actually obtained a standard score of .00 (mean score) on the Mechanical Comprehension test, the part contributed by his +1.5 standard score on the Adaptability test could be roughly corrected for by substracting .50 × (+1.5) from his Mechanical Comprehension standard score, thereby assigning him a corrected standard score on the latter test of − .75.

The above correction principle may be expressed by the following formula providing that all scores are converted into standard score units: Corrected Special Ability Score = Special Ability Score $-r \times$ General Ability Score.

The effect of using this correction formula is to raise the special ability score of an individual who is below the mean on the general ability test and to lower this score for an individual who is above the mean on the general ability test. The raising or lowering is of an amount proportional to the relationship between the two tests and the deviation from the mean on the general ability test. This correction formula serves to reduce the correlation between general ability test scores and corrected special ability test scores to zero, thereby eliminating the variance contributed by so-called general intelligence from a test designed to measure special ability or aptitude. Application of the correction would tend to obviate such disconcerting phenomena as bright but mechanically inept individuals scoring high on the Army Mechanical Ability Test and dull but mechanically gifted garage mechanics scoring low on this test by virtue of the aggravatingly high relationship between scores on the Mechanical Aptitude Test and the General Classification Test.

Incidentally, if corrected scores are to be computed for a large number of individuals, considerable economy may be effected by constructing tables which will enable one to read the corrected scores in either raw or standard score form directly from the table. For any given r a table can be easily constructed in which the special ability test scores are arranged in progression along the vertical and the general ability test scores along the horizontal, or vice versa, and the corrected scores found at the point of intersection in the table.

Corrected scores should be used only if the following conditions are fulfilled:

1. When the special ability test being considered is substantially correlated with a standard test of general intelligence for a particular sample. Otherwise, little is gained by introducing a correction factor which would necessarily be rather insignificant.

2. When success in an occupation for which the special ability test score is used as a predictor is not related to general intelligence. Otherwise, one would lose by eliminating the effect of a factor which is positively related to success.

3. Where there is an empirically demonstrated relationship between corrected scores and occupational success in excess of simply using the uncorrected special ability test scores. Since there is more work involved in obtaining a corrected score, it should justify itself by adding to the predictive efficiency for success in a given job. This is a crucial point since the rationale for the correction is primarily a practical one.

It is anticipated that corrected special ability test scores will find their greatest usefulness in the prediction of success in the semi-skilled trades and possibly in routine clerical jobs.

The Rorschach Test in Industrial Selection

Audrey F. Rieger

Robert N. McMurry & Co., Chicago, Illinois

The place of the Rorschach inkblot test in clinical work has been well established, and some claim made that it is of value in vocational guidance. Another field to which the test can contribute valuable information is that of selection of industrial personnel.

The problem in selection is the choice of a worker who can fulfill the requirements of experience and ability for the job and who is a good risk for long-term employment. He must have the necessary skills and also be able to fit into the organization. Information about his adaptability to the job and to the company is very difficult to get, although some of it may be learned from interviews, references, and tests.

Personality questionnaires have often been used as aids in selection. The applicant, however, is frequently able to tell what the answers imply and finds it to his advantage to falsify his responses, if necessary, to give the impression he believes is favored for the position. The use of validating keys may permit detection of falsification, but they give little idea of the direction of the distortion.

A projective technique such as the Rorschach test makes falsification impossible, since the applicant can in no way determine what the examiner is looking for. The applicant must interpret the unstructured stimuli of the test in his own manner and is unable to determine how to produce a desired picture.

Projective techniques have some disadvantages, however. They are usually time-consuming and always require careful interpretation, a process which demands long and careful training. Therefore the cost of administering these techniques may make it impractical to use them, except for jobs which involve at least a moderate investment on the part of the employer.

As a result, few companies, with the exception of large organizations, can afford to add a Rorschach worker to their staffs. Providing them with access to the services of one on a consultant basis makes it possible for these employers to have the benefit of some information about the personality of the applicant when it is desirable and to pay for it only as it is needed.

For this reason, the services of a Rorschach worker were made available to the clients of a firm of personnel consultants. The test was

always given in conjunction with the regular selection technique, which
is based on a Patterned Interview procedure.[1] Processing of the appli-
cant included paper-and-pencil tests, the interview, and the Rorschach.
The data derived from these sources were then weighted in order to make
a recommendation to the employer with regard to the applicant's chances
to be successful on the job.

In most instances the Rorschach test is of more value if given in ad-
vance of the interview; a brief report can then be made to the interviewer,
with emphasis on clues which he can follow up in the interview.   Oc-
casionally, however, it was not possible to give the test prior to the inter-
view.   At such times, ratings for the job could be assigned independently
from the interview findings and from the test results.

Table 1
Occupations of Subjects

Occupations	Frequency
Personnel Assistant	7
Personnel Director	4
Office Work	4
Sales	3
Engineer	2
Production Manager	2
Industrial Engineer	2
Merchandising Trainee	2
Market Research	2
Production Assistant	1
Reporter	1

Ratings based on the Rorschach results alone involved a comparison
of the strengths and weaknesses reflected in the test results with the
specific requirements of the job.   For example, an applicant for an
executive position was considered less promising if his record indicated
difficulty in organizing abstract material or in controlling his impulses,
whereas one who showed strength in these areas was more likely to be
given a more favorable rating.

Under the special conditions of independent ratings, a total of thirty
applicants were studied.   Table 1 shows the occupations represented in
the group.

Ratings for each of these subjects were made by the interviewer and
by the Rorschach worker with the specific job in mind. Table 2 gives
the distribution of the ratings, from 1 (superior) through 4 (reject).

[1] R. N. McMurry, *Handling personnel adjustment in industry.* New York: Harper,
1944, pp. 297 + xi.

The coefficient of correlation between the two sets of ratings is $+.75 \pm .05$, a very significant result. A correlation of this magnitude indicates that use of either procedure will agree quite well with the results of the other. That the interview alone has great predictive value has previously been shown.[2] Using it as the criterion, it can be assumed that the results of the Rorschach are also valuable for industrial prediction.

It must be noted, however, that after only a brief period of using the test a correlation of this magnitude probably could not be achieved. The interviewer and the Rorschach worker had been associated over a relatively long period of time and were both familiar with the factors on which the recommendations were based and the methods used by the other in weighting these factors. This undoubtedly tended to raise the correlation.

Table 2
Scatter Table Showing Relation Between Rorschach Ratings
and Interviewer's Ratings of Job Applicants

Rorschach Ratings	Interview Ratings			
	4	3	2	1
1	..	..	..	..
2	..	1	11	..
3	1	10	3	..
4	2	2	..	..

Nevertheless, it must be recognized that use of the Rorschach test by itself, a practice which is not recommended under any circumstances, would have led to selection of the best candidate in many more cases than chance alone would suggest. Such results would be found, however, only in instances where the Rorschach worker knew his instrument well.

The most efficient use of the technique is as a supplement to other employment procedures. It should at no time supplant them, as it cannot always assess properly the importance of various personality factors. In addition, it gives no information about the skills the individual possesses, his motivation to work, and several other factors which influence job success. Furthermore, unless care is used, the interpretation may be influenced by the bias of the examiner.

At the same time, the Rorschach offers unique help in learning many facts about job applicants (particularly at the higher occupational levels) and thus aids in improving prediction. The Rorschach test can be used objectively and offers much valuable information for selection.

[2] R. N. McMurry, Validating the patterned interview. *Personnel*, 1947, 24, 263–272.

The Rorschach Test and Occupational Personalities

Audrey F. Rieger

Robert N. McMurry & Co., Chicago, Illinois

The question of differences in personality which may differentiate between occupational groups (and therefore aid in the selection of employees) has been raised, and some attempts have been made to answer it. Dodge, for example (3, 4, 5, 6) found sales and clerical personnel had different patterns of scores on the Bernreuter. Paterson and Darley (10), using the same instrument, were unable to detect differences in their subjects. Verniaud (14), studying saleswomen, clerical workers, and optical workers, reports some differences in MMPI scores which she says correspond with differences in the occupational requirements.

Kaback (8), using the group Rorschach method, noted some statistically significant differences between pharmacists and accountants. Nevertheless, she concluded that neither group showed any generalized characteristics. A less exhaustive study using the same technique, that of Harrower and Cox (7), reports some differences between other occupational groups. Steiner (13) has summarized the Rorschach literature reporting studies of occupational groups.

The present investigation was designed to study personality patterns of certain specific occupational groups as reflected in the individual Rorschach test to determine if differences between such groups do occur and if the differences are meaningful in practical situations.

Subjects

The opportunity to make this investigation into occupational differences in personality arose in the course of routine procedures in the offices of an organization of personnel consultants. Applicants for positions with client organizations are interviewed (9) by one of the consultants and are given such paper-and-pencil tests as seem applicable to the positions for which they are being considered. In addition, the applicants are given an individual Rorschach test by the writer. The Rorschach was adopted for routine use in the employee evaluation program to give the interviewer a fairly objective portrait of the personality of the candidate and to aid in the evaluation of the information elicited in the interview, telephone checks with previous employers, and tests. On the basis of these data, the candidate is then rated with regard to his potential value as an employee.

As a rule, the candidates interviewed have previously been screened by the employer. The men doing this preliminary screening have been trained to be alert to clues indicating instability, inadequate intelligence, and other factors affecting success on the job. As a result, it seems likely that the candidates, most of whom would have been hired if expert advice had not been available, represent a better-than-average group of workers.

Hence these subjects cannot be considered as representative of applicants in their respective occupational fields. Moreover, the occupational classifications, based on the employers' job descriptions, may appear to be somewhat arbitrary, as some of the applicants lacked experience in the field. Since they were believed to have possibilities for the job, however, it seemed reasonable to include them as subjects for study. Some differences which might have occurred between more clear-cut occupational categories may have been obscured as a result, a fact which should not be overlooked in assessing the results of this study.

Table 1
Occupational Classification of Subjects

	N	Age		Years of Experience	
		Range	Mean	Range	Mean
Sales (technical)	55	19–48	29	0–20	3
Engineers	53	21–50	28	0–20	4
Supervisors, foremen	36	23–56	35	0–10	3
Administrators	64	24–48	35	0–18	5
Clerical workers	66	17–45	27	0–20	3
Personnel workers	24	22–45	32	0–11	2
Merchandising trainees	32	20–36	26	....	..
Miscellaneous	22	21–46	30	....	..

The jobs for which the applicants were being considered form the basis for classification into occupational groups. These jobs fall into six categories, with two additional miscellaneous groups. Table 1 summarizes some information about the groups. "Years of experience" noted in the table refers only to experience for the specific job, rather than years of work experience in general.

The supervisors were not being considered for eventual promotion into white collar jobs; they were men who had done well in the shop and were moving up. They had less formal education than the other subjects and did relatively less well on verbal tests, making their best scores on non-verbal items. The clerical group includes accountants, statisticians, and others doing similarly complex work. The last two groups,

the trainees and the miscellaneous, lack homogeneity as they include a wide range of occupations.

Methods

It was not possible to use as subjects only those who were recommended for employment, as the ratings were based on the results of the Rorschach test as well as on the information from the interview and test procedure and could not serve as criteria. Furthermore, the other data about the applicants were not uniform, the interviewing having been done by different individuals, the applicants having taken different tests, etc. As a result, none of the data except the Rorschach test scores could be used.

A large number of Rorschach scores were tabulated by occupational groups. Included were all those given weight in the orthodox interpretation of the results (1), such as color responses, the approach type, etc. In addition, the literature was reviewed for statements about occupational differences in personality which might be represented by Rorschach components; these were tallied. Finally, other scores which on an a priori basis may reflect differences between the groups were also studied.

Table 2 presents the means and standard deviations for all groups for the more important scores studied.[1]

In reviewing the results, the statistical reliability of the scores must be considered. Group differences may be minimized or completely obscured by lack of dependability of the measures. Unfortunately, no conclusive evidence has been put forth in the literature to answer this problem. In general, it is probably true that some of the Rorschach scores possess a high degree of reliability, and others are less dependable. It must also be recognized that ratios and difference scores, which represent relationships between imperfect measures, are less reliable than the component parts. Differences between groups might be hidden by such unreliability, and results with these scores must be taken with caution. Chief among these are $FC-(C+CF)$, $W\%$, etc.

Two statistical methods were used. The first involved testing the differences between the means of the groups for each score to determine if any of the differences were significant (CR at least 3). If such differences occurred consistently, it was planned to make up a composite picture of the worker in each field based on the means of the scores. Although this procedure is contrary to the basic idea of the interdependence of all behavior in the Rorschach test, it had had at least a limited success

[1] Table 2 may be ordered as Document 2552 from American Documentation Institute, 1719 N Street, N. W., Washington 6, D. C., remitting $0.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or $0.50 for photocopies (6 × 8 inches) readable without optical aid.

elsewhere (11) and might aid in the development of occupational personality patterns to be adapted to selection and guidance.

The second method involved chi square tests of a number of scores which are more meaningful if interpreted in relation to each other, an approach difficult to manage in a study such as this where the clinical implications of many of the results must be overlooked for lack of information and an inability to deal with large numbers of subjects on an individual basis.

To equate one set of scores with another, a number of the measures were transmuted into a normalized scale of standard or T-scores, with identical means and standard deviations (2). A T-score on any one scale has the same relationship to the distribution of those scores in the total group as does the same T-score value on any other scale. The scores so treated were chiefly those making up the Approach Type (W, D, and Dd) and the Experience Balance (M and C).

Results

The attempt to find personality differences between the occupational groups had some significant results, but the Rorschach scores, as tested here, would fail to differentiate the groups in practice. Most of the differences appear to be related to variations in response total; for example, if R is high, W, D, or Dd is necessarily high in relation to the scores of other subjects with low R. Whether differences which are dependent on variations in R can be considered as real differences is a question which requires further study.

Tables 3, 4, and 5 summarize the significant differences between means.[2] The chi square tests support some of the results but fail to add much new information.

Only two groups of subjects tend to stand out from the remainder. These are the administrative group and the supervisors and foremen.

The administrators are characterized chiefly by their facility in producing and handling ideas (high R, low A%, etc.). Scores of this group indicate complexity of structure as well as lability and freedom of expression. Most of the areas in which these subjects differ from the remainder of the groups appear to be dependent on superior verbal facility, however, since the significance of the differences disappears when R is taken into account.

The supervisors and foremen form a fairly homogeneous classification and appear to be truly different from the other subjects. The group of supervisors, however, would probably stand out less noticeably if com-

---

[2] For brevity, the tables are omitted. The writer will be glad to supply the information upon request.

pared with similar workers; this is the only group of subjects not of the white collar or professional classes, a fact which must not be overlooked in reviewing their scores.

The supervisors are characterized by: limitation of ideas (low R), narrow range of interests (high A%), rigidity in judgment (high F + %), and restricted emotional life (low M and C). They are ill at ease in close associations with others (H-Hd low), and they tend to avoid contacts with others, even superficial relationships (H% low).

These restrictions may be explained by a number of factors. Probably most important is the fact that these men work chiefly with their hands and rarely deal with verbal concepts. Their weakness in verbal matters is evidenced by the relatively poor showing in verbal tests. In addition, the relatively impoverished background must be noted, with emphasis on the lower level of education. Finally there is the possibility that the personality may be reflected in the choice of occupation.

Although some traits seem to differentiate between the other groups (i.e., the salesmen seem to be more concerned with problems of health than are the other groups), these differences in scores may result from chance factors in the selection of the applicants rather than from specific group differences. The chief exception to this is the suggestion that the engineers show less interest in other people (H% low), a finding which corroborates the results of many other studies. In the case of the administrators also, the personality characteristics noted here can be related to some extent to the requirements of the job.

Discussion

The fact that the individual Rorschach test reflected variations between the occupational groups suggests that the Rorschach is sensitive to differences between groups. The test probably lacks reliability when used statistically to study differences between groups, however, and may therefore hide or distort some actual differences.

Another limitation is the selection of above-average rather than representative subjects. These may be more similar than would be subjects chosen at random.

Furthermore, the occupational classifications used in this study were not homogeneous categories of jobs. Although they were based on the employers' job titles, the duties under each heading varied widely. "Salesman," for example, might mean one who sold machinery or one who merely created good will for his employer's products. As a result, the heterogeneity of the jobs within each classification probably lessened the chances of turning up significant differences.

The fact that so few occupations could be differentiated in this in-

vestigation is of significance for selection and guidance. It is unlikely that the technique is wholly at fault in not turning up differences, since some consistent differences were noted. In some rare instances, such as cases where the individual possesses some special talent, the choice of occupation is determined by the talent. It is more likely that the occupation in which the worker spends most of his life is almost a matter of chance, determined by opportunity, rather than an end result directed by specific personality structure.

If this is true, no occupation can be said to draw people of similar personality makeup, although it may influence them to the extent that they later appear similar. There are a few exceptions to this, such as certain research fields (12). In general, however, any single personality pattern can be fitted into a number of jobs which may appear to differ greatly in demands on the individual; this is indicated by the great overlapping of scores between groups, even when the means differed significantly. No single personality type can be associated with any of the occupational groups, nor can it be assumed that any particular type of personality occurs to excess in any occupational group.

Recommendations for hiring must be based not on a general pattern for an occupation but on the specific requirements of the job and its place within a functioning organization. Here the Rorschach test can be of great value in pointing out the applicant's strengths and weaknesses, with due consideration to the part he will play in the particular organization and without concern for a generalized occupational pattern.

Summary

A study of several occupational groups by means of the individual Rorschach test showed a few statistically significant differences between groups. The only important result is the distinction found between those who deal with verbal concepts (chiefly administrators but including salesmen, engineers, clerical workers, and personnel workers) and those who work with their hands (supervisors and foremen). Personality patterns cannot be reliably used for placement, selection, and guidance. These findings (the lack of patterns) should not be construed as a denial of the importance of the descriptive elements of the Rorschach results in selection and guidance.

References

1. Beck, S. J. *Rorschach's test.* New York: Grune & Stratton, 1944–1945, 2 vol.
2. Cronbach, L. J. A statistical method for treatment of limited patterns of scores. Unpublished MS, Univ. Chicago, 1948.
3. Dodge, A. F. Social dominance and sales personality. *J. appl. Psychol.,* 1938, 22, 132–135.

4. Dodge, A. F. What are the personality traits of the successful sales person? *J. appl. Psychol.*, 1938, 22, 229–238.

5. Dodge, A. F. What are the personality traits of the successful clerical worker? *J. appl. Psychol.*, 1940, 24, 576–586.

6. Dodge, A. F. Characteristics of good clerks. *Person. J.*, 1942, 20, 324–327.

7. Harrower, G. J., and Cox, K. J. The results obtained from a number of occupational groupings on the professional level with the Rorschach group method. *Bull. Can. Psychol. Assn.*, 1942, 2, 31–33.

8. Kaback, Goldie Ruth. *Vocational personalities.* Teach. Coll. Contr. Educ., No. 924. New York: Columbia Univ. Press, 1946.

9. McMurry, R. N. *Handling personnel adjustment in industry.* New York: Harper, 1944.

10. Paterson, D. G., and Darley, J. *Men, women, and jobs.* Minneapolis: Univ. Minn. Press, 1936.

11. Rieger, Audrey. Rorschach analysis of adolescent groups. Unpublished MS, Univ. Chicago, 1945.

12. Roe, Anne. Rorschach study of a group of scientists and technicians. *J. consult. Psychol.*, 1946, 10, 317–327.

13. Steiner, Matilda E. The use of the Rorschach method in industry. *Rorschach Res. Exch.*, 1947, 11, 46–52.

14. Verniaud, Willie Maud. Occupational differences in the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1946, 30, 604–613.

A Factorial Analysis of Arm–Hand Precision Tests *

Robert H. Seashore, Frank J. Dudek, and Wayne Holtzman

Northwestern University

Steadiness and precision of movement have long been thought to be a significant variable in various kinds of motor performance. Emphasis had been placed on the prediction of various skills by the use of various psychomotor tests and in the work done thus far there has been a tendency to accept a general factor or component of "steadiness." Seashore (3) sums up the conditions necessary for such a factor as follows: "According to the hypothesis of a group factor for steadiness, those coordinations which emphasize accuracy (precision or steadiness) while minimizing speed and strength should cluster together. Such steadiness tests, in wide variety, should intercorrelate moderately or highly, and show no correlations with speed and strength tests."

Various studies have presented evidence supporting the hypothesis of a group "steadiness" factor. Spaeth and Dunham (5) working with 73 army men, studied the relationship between Dunlap's test of precision in thrusting (1) and rifle target shooting. The correlation between the two tests for subjects ranging from poor to expert marksmen was .61, a very significant relationship. Seashore and Adams (4) found test intercorrelations of .45 or higher with a battery of five steadiness tests (postural sway with eyes closed, rifle muzzle sway when sighting, hand tremor, stylus thrusting at holes, and stylus held stationary in holes). Humphreys, Buxton, and Taylor (2) reported intercorrelations ranging from .37 to .69 with a median coefficient of between .52 and .55 between thrusting steadiness, stationary steadiness, an ataxiameter, and rifle sway. Relatively little research has been reported concerning the nature of the factors responsible for differences in performance on various tests measuring precision of movement and steadiness.

Purpose of Study

The present study was an attempt to determine the nature of factors underlying performance on seven measures of visuo-motor co-ordination.

These measures emphasized precision of movement of the preferred arm and hand. The tests used in this battery were selected with several considerations in mind: 1. they should be relatively uninfluenced by strength and speed; 2. they should be relatively free from the effects of muscular fatigue; 3. it should not be possible to get a high score by "trick" performances, and, 4. there should be little practice effect.

Tests Used

Seven tests were selected for inclusion in the battery. It is recognized that these seven tests do not sample, in all probability, the entire range of steadiness measures. However, it was not possible to include more variables in this battery because of time limitations. The more promising tests as determined from the analysis to be described will be included in another battery in an attempt to study and define more completely the domain of "steadiness." The seven tests are described below:

1 and 2. The Universal Ataxiameter: This test was designed to measure the horizontal and vertical components of involuntary movement of the hand and forearm. The subject attempted to hold a wooden tab or handle as motionless as possible. Movements were magnified by means of a leverage system and photo-electric cells recorded the amount of movement made. Horizontal and vertical components were scored separately. Five trials of 15 seconds each were administered in each cycle of tests.

3. Seashore Photoelectric Target Register (Revised): This test emphasized aiming and constant adjustment of a circular beam of light to a target. From a mirror which was on the end of a rod controlled by the subject the beam of light was reflected into a small hole. The beam of light activated a photo-electric cell which recorded the time the individual was "on the target." If the aim was perfect (i.e., the circle of light completely covering the hole) the counter recorded 10 counts per second—if the beam was only partly on the target the counts per second were correspondingly less. Five trials of 15 seconds each were administered in each cycle.

4. Straight Tracing Test: This test was a modification of the V-slot tracing test described by Whipple. At a controlled speed the subject drew a wire stylus between two brass plates without any base plate. The path formed by the brass plates was a converging one and the direction of the hand-movement was toward the body. If the stylus touched either side of the path it activated a very sensitive Potter Electronic Counter which counted at a rate of 60/sec. during time the stylus was in contact. Five trials were administered in each cycle.

5. Curved Tracing Test: This test was a variant of the straight trace. The path was of the same width throughout, but it was irregularly curved. The subject moved a wire stylus from left to right along this curved path. Scores were obtained on the Potter Electronic Counter as mentioned above. Ten trials of 15 seconds each were administered.

6. Sine-Curve Rod Tracing Test: In this test the subject moved a ring stylus of $\frac{7}{16}$ inch inside diameter along a brass rod of $\frac{3}{8}$ inch diameter. The rod was bent vertically in the form of a sine curve. The direction of arm-movement was from left to right. Time of contact between the ring and the rod was recorded by the Potter Electronic Counter. Speed of movement was controlled so that each trial required about 30 seconds. Five trials were administered per cycle.

7. *Three-dimensional Rod Tracing Test:* This test was constructed and scored similarly to the previous test, but the rod was bent into an irregular shape in three dimensions. The ring stylus had a somewhat larger inside diameter (1 and ⅛ inches) to permit greater freedom of movement. Five trials requiring approximately 40 seconds each were administered during each cycle.

8. *Thrusting Steadiness Test:* This test was a modification of Dunlap's test used by Seashore, Adams and others. The subject thrust a stylus into a hole in time with a metronome at a rate of one thrust every two seconds. Holes of three diameters were used, with S making 10 thrusts into each size hole in five different trials. The score was the number of thrusts made without contacting the side of the hole.

Method

This battery of tests was administered under a cycle plan. On one day the subject went through two successive cycles. Each cycle of tests required approximately 40 minutes to administer. Forty-eight hours later the subject returned and repeated two more cycles of the complete battery. In this way it was possible to obtain measures of reliability during any one testing period and for test-retest periods. Reliabilities are indicated in the last three columns of Table 1. The first two columns contain uncorrected, test-retest reliabilities for cycles combined in various ways. The last column is an estimate, corrected by the Spearman-Brown formula, of the reliability of the total score (i.e., all four cycles). The test-retest reliabilities range from .54 for the horizontal component of the ataxiameter to .85 for the three-dimensional rod test. These co-

Table 1

Intercorrelations, Means, S.D.s, and Reliabilities of Tests in
Steadiness Battery (As computed for Group II, N = 100.)

Variable	No.	Product-moment r's.							M	S.D.	Cycles 1+2 vs 3+4	Cycles 1+3 vs 2+4	Of All Four Cycles*
		1	2	3	4	5	6	7	M	S.D.			
Atax. (Hor.)	1								5.19	2.91	54	67	76
Atax. (Ver.)	2	47							4.43	2.68	75	84	89
Targ. Regis.	3	44	44						5.95	2.65	79	83	89
Str. Trace	4	39	20	40					3.10	2.76	79	82	89
Curv. Trace	5	30	04	26	61				4.77	3.47	82	95	95
Sine Trace	6	31	−06	26	57	72			4.43	3.12	77	80	91
3-dim. Trace	7	22	−07	34	37	72	83		4.08	3.68	85	82	91
Thrust	8	31	35	43	66	38	41	49	4.03	2.64	78	88	91

* Corrected according to the Spearman-Brown formula.

efficients indicate a rather high degree of stability for day-to-day measurement of precision of movement and steadiness.[1]

Results

The tests in this battery were administered to 100 volunteer subjects selected from elementary psychology classes at Northwestern University. All were right-handed males. Since stability of day-to-day measures seemed fairly high the scores used for analysis were the sums of scores on all performances for any given test. Product-moment correlations were then computed for these scores. These intercorrelations are presented in Table 1. The variables are, in general, positively correlated. No correlations are significantly negative though several are not significantly greater than zero.

Table 2

Centroid and Rotated Factor Loadings of Tests in Steadiness Battery

| | | Group II (N = 100) | | | | | | | Group I (N = 39) | | | |
| | | Centroid Loadings | | | | | Rotated Loadings | | Rotated Loadings | | | |
Variable		I	II	III	h_c^2	h_v^2	I	II	III	I	II	III	h^2
Atax. (Hor.)	1	56	31	22	49	46	23	27	[58|				
										20*	10*	[77*|	64*
Atax. (Ver.)	2	37	58	24	54	53	−06	18	70|				
Targ. Regis.	3	60	32	25	53	52	25	27	[62|	22	03	[77]	63
Str. Trace	4	78	09	−50	83	86	23	[88]	10	−03	[66]	11	45
Curv. Trace	5	72	−40	−09	68	68	[68]	46	−03	27	45	46	48
Sine Trace	6	74	−53	−03	83	84	[81]	41	−08	[84]	20	29	83
3-dim. Trace	7	72	−56	17	85	86	[89]	24	00	[83]	17	14	74
Thrust	8	69	17	−17	55	53	25	[61]	31	19	[66]	30	57

* In the data for Group I the horizontal and vertical scores for the Ataxiameter were combined into a single score.

The matrix of intercorrelations was subjected to a centroid factorial analysis. The results of this analysis are presented in Table 2. Most significant here are the two matrices of rotated factor loadings. The data for Group I were obtained in a preliminary investigation carried out by Holtzman in the summer of 1946. Only 39 individuals were used as subjects for this study. Nevertheless, the intercorrelations resulting from this sample were factorially analyzed with the resulting factors shown in

[1] Note that the coefficients of reliability which compare trials 1 and 2 of the first day with trials 3 and 4 of the second day are somewhat lower than the correlations between trials 1 and 3 vs. trials 2 and 4, which tend to balance out the diurnal vairation. G. Paulsen, *J. appl. Psychol.*, 1935, Vol. 19, pp. 166–79 and pp. 29–42 found, as we do, that there was a lower intercorrelation between trials on successive days than on the same day, on one test of hand steadiness. However there was still a very significant correlation between days.

the right-hand part of the table.   In the present study data were gathered for 100 individuals (designated Group II).   It is noteworthy that the factorial picture is so similar from group to group.   Major factor loadings in each sample have been enclosed in boxes.   It is apparent that the factor structure is practically identical from sample to sample.

Three factors were adequate to account for the correlations in each sample.   When these were rotated with criteria of simple structure and positive manifold in mind the factors resulting from each analysis could be identified as similar.   These factors were identified and named as follows:

Factor I has primary loadings on the three-dimensional rod test and on the sine-curve rod test.   It has a major loading as well for the curved tracing test for Group II, but not for Group I.   This factor seems obviously to be associated with steadiness and precision of movement which involves spatial components in two or more planes.

Factor II has major loadings in the straight tracing test and in the thrusting steadiness test.   It will be recalled that in the straight tracing test the subject moved a stylus along a straight path toward his body. In the thrusting test the subject thrust a stylus toward a hole away from his body, but in a very similar path.   This factor, it would seem, is associated with precision of movement in a restricted plane.

Factor III has major loadings for the target register and for the ataxiameter scores.   The common component here is involuntary movement of the arm and hand.   This factor seems to be what is usually thought of as "steadiness."

An examination of the factor loadings suggests that simple structure has not been ideally achieved and that the factors may be somewhat correlated.   Correlations among the factors were estimated graphically by determining the cosine of the angular separation between oblique vectors representing them.   Factors I and II are moderately correlated.   Factor III does not seem to be highly correlated with the other two.   It would seem, then, that stationary steadiness is relatively independent of steadiness where movement is involved.   The correlation between factors I and II should not be surprising, since both involve movement.   The number of dimensions within which movement takes place seems the important difference.   Factor I included those tests calling for movement in a three-dimensional space, while Factor II was restricted more or less to movements in a restricted plane—or two-dimensional space.   There is another possible interpretation based on the amount of movement involved. Those tests involved in Factor II require relatively shorter movements than do those appearing in Factor I.   (The only anomaly here is the curved trace test which in this respect is more like the tests appearing on Factor II.)   In this way it might be that Factor I represents tests in-

volving a general bodily orientation and control whereas Factor II tests represent finer adjustments within but one member of the body (the hand and arm). This hypothesis remains to be investigated.

Summary

The results of this investigation suggest several important considerations.

1. The intercorrelations among various tests all of which presumably depend on steadiness and precision of movement cannot be adequately accounted for by postulating a single factor of steadiness. There are, it would seem, various components influencing performance on these seemingly similar kinds of tasks.

2. There may be yet other factors accounting for scores made on steadiness tests measuring only involuntary kinds of arm and hand movement. This is suggested by the differences between the common factor variance and the reliabilities of the tests measuring involuntary movement in the present battery. While the reliabilities would indicate from 80 to 90% of the variance accounted for, common factors seem to account only for about 50% of the test variance. This would mean that about one-third of the variance is attributable to some specific factor or to a common factor yet to be identified by including these tests in another battery with other types of tests.

3. The results indicate that stationary steadiness (or involuntary movement of the arm and hand) is not highly related to precision of movement. Those factors, however, which involve spatial components are related to a greater degree.

These results and conclusions have implications in the area of selection in certain industrial areas. Many kinds of tasks are recognized as depending upon "precision," "steadiness," or "coordination." However, it would seem that there are several identifiable components—the isolation of which would improve our ability to select for the particular components which are represented in particular jobs. Thus—a lathe operator and drill-press operator might not require the same kinds of "steadiness."

References

1. Dunlap, K. Improved forms of steadiness tester and tapping plate, *J. exp. Psychol.*, 1921, **4**, 430–433.

2. Humphreys, L. G., Buxton, C. E., and Taylor, H. R. Steadiness and rifle marksmanship, *J. appl. Psychol.*, 1936, **20**, 680–688.

3. Seashore, R. H. An experimental and theoretical analysis of fine motor skills, *Amer. J. Psychol.*, 1940, **53**, 86–98.

4. Seashore, R. H., and Adams, C. R. The measurement of steadiness: a new apparatus and results in marksmanship, *Science*, 1933, **78**, 285–287.

5. Spaeth, R. A., and Dunham, G. C. The correlation between motor control and rifle shooting, *Amer. J. Physiol.*, 1921, **56**, 249–256.

What Do Readership Studies Really Prove? *

H. P. Longstaff and G. P. Laybourn

University of Minnesota

Almost every publication seems to have "just made a reader study" on the basis of which it seems to be able to prove that it "leads the field in readership." Such violently conflicting claims have become so common that many a reader, advertiser, and publisher—confused by this contradiction . of alleged facts—has begun to ask: "What do readership studies really prove?" One such perplexed publisher set out to try to answer this question.

Purpose

The purpose of the study conducted by the Putman Publishing Company was to call attention to fallacies inherent in "readership studies" as they have been commonly conducted and to suggest the need for more careful scrutiny of the results of such studies. It was *not* the purpose of this investigation to set up an entirely new, flawless technique for the study of relative readership. This was a purely analytical investigation of readership-study methods and techniques.

Procedure

To appraise the validity of the "orthodox" type of readership study, a procedure was devised to compare the relative readership standings of three industrial magazines on the basis of three different readership-study techniques which yielded, respectively, (1) the number of readers based upon the number of "mentions" [1] obtained in response to an original questionnaire employing "orthodox" readership-study techniques, (2) the number of readers corrected for "votes" [2] obtained in response to a follow-up questionnaire, and (3) the number of readers corrected for

* This paper is a condensation and revision of an investigation conducted by Mr. R. L. Putman of the Putman Publishing Company, and is published with their permission. The original study entitled "We Made a Reader Survey" was published by the above mentioned company. Copies of the original report are available (without charge to industrial advertisers and their advertising agencies) upon application to the Research Department, Putman Publishing Co., 737 North Michigan Avenue, Chicago 11, Illinois.

[1] *Mention* refers to the naming of a magazine in reply to the question, "What magazines do you read?"

[2] *Vote* refers to a "YES" response to the question, "Do you read this magazine?"

votes and "disqualifying negative comments" [3] obtained in response to the follow-up questionnaire.

This procedure entailed the use of three different questionnaires: (1) an original "orthodox" type of questionnaire, sent to 1,000 known readers of Magazine *A*, asking "What magazines do you read?"; (2) a follow-up questionnaire, sent to those who failed to mention Magazine *A* in response to the original questionnaire, asking "Do you read Magazine *A*?" and providing for comments on Magazine *A*; and (3) a follow-up questionnaire, sent to those who failed to mention Magazine *B* and/or Magazine *C* in response to the original questionnaire, asking either "Do you read Magazine *B*?" or "Do you read Magazine *C*?" and providing for comments on either Magazine *B* or Magazine *C*.

The First Questionnaire. The original questionnaire appeared on the letterhead of an independent manufacturer, who supposedly was attempting to determine in what industrial magazines he should place some advertising, and requested the addressee to write down the names of the industrial, business, or trade magazines which he reads. This first questionnaire was mailed to 1,000 "known readers" of Magazine *A* whose names and addresses had been copied from response slips[4] which readers[5] had taken from issues of Magazine *A*. Thus, Questionnaire No. 1 provided a check of what known readers say they read, after they had proved their readership of one publication, without knowing that the questioner knew anything about what they had read.

The Second Questionnaire. Since it was found that over one-half of the "known readers" of Magazine *A* who replied to Questionnaire No. 1 failed to mention Magazine *A*, it was decided to send them a follow-up questionnaire in order to discover why they had failed to do so. This second questionnaire appeared on the letterhead of the same manufacturer, who supposedly wondered whether the addressee's failure to mention Magazine *A* in replying to the first questionnaire was merely an oversight, and requested the addressee to indicate whether or not he read this magazine and provided a space for him to comment upon it. Thus, Questionnaire No. 2 was sent to those persons who, in replying to the original questionnaire, had failed to mention Magazine *A*.

---

[3] *Disqualifying negative comment* refers to a comment which indicates that a respondent who replied "YES" to the question, "Do you read this magazine?" actually does not read it.

[4] These slips had been inserted into copies of Magazine *A*; readers filled in subjects on which they wished more information, signed their names, and mailed the slips to the publisher.

[5] These readers were not necessarily subscribers to Magazine *A*: in several cases it was apparent that these inquirers had sent slips taken from copies of the magazine received by someone else.

The Third Questionnaire. Since it was found that nine out of every ten respondents replying to Questionnaire No. 2 reported that they read Magazine *A* (despite their failure to mention Magazine *A* in response to Questionnaire No. 1), it was decided to duplicate as closely as possible the conditions under which the second questionnaire had been sent out asking about Magazine *A* by sending out a third questionnaire asking about Magazine *B* and Magazine *C*. In order that no one would receive more than one follow-up questionnaire, Questionnaire No. 3 was not sent to those who had been sent Questionnaire No. 2, i.e., those who had failed to mention Magazine *A*. Thus, the third questionnaire was sent to those who, in replying to the original questionnaire, had failed to mention Magazine *B* and/or Magazine *C* but had mentioned Magazine *A*. Questionnaire No. 3 duplicated the conditions of Questionnaire No. 2 as closely as possible with the letter as nearly as identical as possible. Each questionnaire dealt with only one of the two magazines, Magazine *B* or Magazine *C*, and requested the addressee to indicate whether or not he read this magazine and provided a space for him to comment upon it.

Results and Discussion

The results of this investigation will be discussed in terms of (1) the response to each of the three questionnaires, (2) the effect of combining the results of the original questionnaire with the results of the two follow-up questionnaires, and (3) the interpretations necessitated by the comments accompanying the replies to the follow-up questionnaires.

The data obtained in response to each of the three questionnaires are presented in Table 1.

The First Questionnaire. In response to the 1,000 letters mailed to "known readers" of Magazine *A*, 585 replies were received,—a 58.5% response. Referring to Table 1, it will be noted that, of these 585 respondents "1,000% salted" for Magazine *A*, only 47.3% mentioned Magazine *A*. Thus, on the basis of the original questionnaire, Magazine *A* ranked third in readership.

The Second Questionnaire. Of the 291 who had failed to mention Magazine *A* in their replies to the first questionnaire and who were addressed with Questionnaire No. 2, 222 replied to the second questionnaire, —a 76.2% response. Referring to Table 1, it will be noticed that, of these 222 respondents, 91.4% replied "YES" to the question, "Do you read Magazine *A*?"

Why did so many of these respondents fail to mention Magazine *A* in replying to the first questionnaire? How could so many "change their minds" in replying to the second questionnaire? Surprisingly, 150 of those replying to Questionnaire No. 2 made some sort of comment —

Table 1

Response to Each of the Three Questionnaires

Original "Orthodox" Questionnaire: "What magazines do you read?"						
	1st Questionnaire					
	Magazine *A*		Magazine *B*		Magazine *C*	
Replies	No.	Per Cent	No.	Per Cent	No.	Per Cent
Mentioning as Read	277	47.3	309	52.8	295	50.4
Not Mentioning as Read	308	52.7	276	47.2	290	49.6
	585	100.0	585	100.0	585	100.0

Follow-up Questionnaires: "Do you read this magazine?"						
	2nd Questionnaire		3rd Questionnaire			
	Magazine *A*		Magazine *B*		Magazine *C*	
Replies	No.	Per Cent	No.	Per Cent	No.	Per Cent
YES	203	91.4	23	41.0	30	52.6
"Occasionally"	5	2.2	6	10.7	5	8.8
NO	13	5.9	27	48.3	21	36.8
No Reply	1	.5	0	0.0	1	1.8
	222	100.0	56	100.0	57	100.0

surprisingly, because such comment, while suggested, was not specifically asked for. This voluntary comment of these readers was perhaps the most revealing part of the response to the second questionnaire. Some typical comments follow:

"I have received Magazine *A* for nine years and have asked for more information numberless times."

"Overlooked this originally—I know of at least 15 other men in our organization who read Magazine *A* regularly."

"Sorry to have overlooked Magazine *A* as this is really one of my favorite magazines along with Magazine *E* and really like it very much."

The Third Questionnaire. Of the 74 who had failed to mention Magazine *B* in their replies to the first questionnaire and who were addressed with Questionnaire No. 3, 56 replied to the third quetsionnaire,— a 75.6% response; and, of the 74 who had failed to mention Magazine *C* in their replies to the first questionnaire and who were addressed with Questionnaire No. 3, 57 replied to the third questionnaire,—a 77.0% response. Again referring to Table 1, it will be observed that, of those replying to Questionnaire No. 3, 41.0% of those who had failed to mention

Magazine B in the first questionnaire replied "YES" to the question, "Do you read Magazine B?"; and 52.6% of those who had failed to mention Magazine C in the first questionnaire replied "YES" to the question, "Do you read Magazine C?"

Effect of Combining Results

We have just seen that the percentages of "YES" replies to the question, "Do you read this magazine?" asked in the follow-up questionnaires, were 91.4, 41.0, and 52.6, respectively, for Magazine A, Magazine B, and Magazine C. Taking these percentages of each publication's "replies-failing-to-mention" in response to the original questionnaire, we note the following:

	Percentage Saying They Read in Response to Follow-up Questionnaire	Number Failing to Mention in Response to Original Questionnaire	Additional Readers
Magazine A	91.4	308	282
Magazine B	41.0	276	113
Magazine C	52.6	290	153

Then, adding the results of the follow-up questionnaires to the results of the original questionnaire, we arrive at the following:

	Number of Readers from Original Questionnaire	Number of Additional Readers from Follow-up Questionnaires	Total Readers
Magazine A	277	282	559
Magazine B	309	113	422
Magazine C	295	153	488

Thus, contrasting the results of the original "orthodox" study with the final figures obtained from all three questionnaires, we note the following inversion in the ranks of the three publications:

	Results from Original Orthodox Study		Figures from All Three Questionnaires	
	Readers	Rank	Readers	Rank
Magazine A	277	3rd	559	1st
Magazine C	295	2nd	448	2nd
Magazine B	309	1st	422	3rd

It should be noted that these final figures are *not* presented as accurate measurements of the relative readerships of these three magazines. Rather, it is believed that these final figures, contrasted with the figures of the first "orthodox" questionnaire, give evidence of the fallacies inherent in such "orthodox" readership studies.

Interpretations Necessitated by the Comments

What is perhaps the most revealing part of this entire investigation is found in the comments which were made by those replying to the follow-up questionnaires as to their readership of, and their opinions of, these publications.[6]

Comparison of Comments of Those Replying to Follow-up Questionnaires. The comments accompanying the replies in response to Questionnaire No. 2 (regarding Magazine *A*) and to Questionnaire No. 3 (regarding Magazine *B* or Magazine *C*) may be categorized as follows: (1) *Favorable* —i.e., the comment definitely reveals the respondent's approval of the publication and/or his actual readership of it; (2) *Negative*—i.e., the comment definitely shows that the respondent either does not actually read the publication regularly or does not feel that it is valuable to him; and (3) *Non-committal*—i.e., the comment tells nothing definite as to the respondent's actual readership or his personal opinion of the magazine's value to him.

Table 2

Comparison of Comments Accompanying Replies
in Response to Follow-up Questionnaires

| | In response to: 2nd Questionnaire | | 3rd Questionnaire | | | |
| | Magazine *A* | | Magazine *B* | | Magazine *C* | |
Replies	No.	Per Cent	No.	Per Cent	No.	Per Cent
With Comments	150	67.5	27	48.2	44	77.1
Without Comments	72	32.5	29	51.8	13	22.9
Total Replies	222	100.0	56	100.0	57	100.0
Comments	No.	Per Cent	No.	Per Cent	No.	Per Cent
Favorable	102	68.0	6	22.2	10	22.7
Negative	11	7.3	18	66.6	29	65.9
Non-committal	37	24.7	3	11.2	5	11.4
Total Comments	150	100.0	27	100.0	44	100.0

Referring to Table 2, it will be seen that Magazine *A* elicited a strikingly larger percentage of "favorable" comments and a strikingly smaller percentage of "negative" comments than did either Mazagine *B* or Magazine *C*. From this tabulation there seems to be strong evidence, first, in the case of Magazine *A*, that actual comments show far greater

[6] A complete tabulation of all comments is presented in the original report, *We made a reader survey*, pp. 22–43.

active readership than mentions on the original questionnaire showed; second, in the cases of both Magazine *B* and Magazine *C*, that actual comments show far lower active readership than both mentions on the original questionnaire and votes on the follow-up questionnaire showed. Thus, it would seem that, in a readership study in which a publication's readership depended upon the number of mentions it receives, there would be a marked tendency for the readership of Magazine *A* to be under-appraised and the readerships of Magazine *B* and Magazine *C* to be over-appraised.

Do "Negative" Comments Disqualify "YES" Votes? An examination of the comments reveals the fact that some of those who voted "YES" on either Questionnaire No. 2 or Questionnaire No. 3 ("Do you read this magazine?") commented to the effect that they really do not read the publication in question. Typical of this type of comment are the following:

> "My recent reading, I am ashamed to say, has been sadly neglected be-cause of a heavy work load. Consequently I have temporarily passed up this magazine. In my opinion it is a good magazine."
> "I read this magazine occasionally but have never inquired about items of interest in its advertisements. It takes too much time to read the magazine and carefully go through its advertisements."
> "Not too helpful in my line of work but have obtained some information from it at various times."

If we accept the principle that such negative comments disqualify the "YES" votes, then, for each of the publications, we may deduct from the number of respondents voting "YES" when asked in the follow-up questionnaire "Do you read this magazine?" the number of such respond-ents whose comments disqualify their "YES" votes, yielding the number and percentage of "YES" votes corrected for "disqualifying negative comments." Following the method of calculation outlined above under "Effect of Combining Results," if we take these corrected percentages of each publication's "replies-failing-to-mention" in response to the original questionnaire, we may obtain the number of additional readers from the follow-up questionnaires corrected for disqualifying negative comments. When these are added to the number of readers obtained in response to the original questionnaire, we arrive at the total number of readers obtained in response to the original questionnaire corrected for both votes and dis-qualifying negative comments in response to the follow-up question-naires. These figures together with the ranks of each of the three magazines are presented in the last two columns of Table 3, which com-pares the relative standings of the three publications on the basis of the three methods of analysis employed.

Table 3

Comparison of the Relative Standings of Three Publications on
the Basis of Three Different Readership-Study Techniques

	Relative Standings on Basis of:					
Publication	Number of Readers Based on "mentions"[1] Obtained in Response to Original Questionnaire Employing "orthodox" Readership-Study Techniques		Number of Readers Corrected for "votes"[2] Obtained in Response to Follow-up Questionnaire		Number of Readers Corrected for Votes and "disqualifying negative comments"[3] Obtained in Response to Follow-up Questionnaire	
	Readers	Rank	Readers	Rank	Readers	Rank
Magazine B	309	1st	422	3rd	398	3rd
Magazine C	295	2nd	448	2nd	412	2nd
Magazine A	277	3rd	559	1st	550	1st

[1] *Mention* = the naming of a magazine in reply to the question, "What magazines do you read?"

[2] *Vote* = "YES" response to the question, "Do you read this magazine?"

[3] *Disqualifying negative comment* = comment which indicates that a respondent who replied "YES" to the question, "Do you read this magazine?" actually does not read it.

It should be noted that none of the figures contained in Table 3 are presented as accurate measurements of the relative readerships of these publications. These contrasting figures, however, do suggest that the further these studies are carried, the greater the discrepancy between the original "orthodox" technique results and the final figures.

What Do Readership Studies Really Prove?

The variations that are revealed in these figures would seem to lay down a challenge to the commonly accepted belief that one can measure readership by asking "What magazines do you read?" Both the "votes" and the comments obtained in the follow-up questionnaires of this investigation reveal the influences of the human tendencies to say we read what we feel we are expected to read, to boast of what we read beyond what we actually do read, and to protect ourselves from possible adverse criticism by stating that we keep up with what we think is accepted as "required reading." It seems apparent, therefore, that asking people "What do you read?" may measure the relative effectiveness of publishers' promotion, publicity, and propaganda over many years, but does not necessarily measure readership.

Summary

1. In order to discover to what extent the results of "orthodox" readership studies are dependable, an attempt was made to compare the

relative readership standings of a number of publications as determined by three different readership-study techniques.

2. In the "orthodox" type of study employing a questionnaire which asked "What magazines do you read?" only 47.3% of those replying mentioned Magazine *A* even though practically everyone to whom the questionnaire had been sent was a "known reader" of Magazine *A*. On the basis of this original questionnaire, Magazine *A* ranked third in readership, Magazine *C* ranked second, and Magazine *B* ranked first.

3. When a follow-up questionnaire was sent to those who, in replying to the original questionnaire, had failed to mention Magazine *A*, Magazine *B*, or Magazine *C*, asking "Do you read this magazine (Magazine *A*, Magazine *B*, or Magazine *C*)?" the relative readership standings obtained in the original questionnaire were reversed.

4. When the comments made on the follow-up questionnaires were taken into account, the readerships of the three publications in question were changed still further.

5. It would seem apparent from this investigation, therefore, that the burden of proof rests on those who conduct "orthodox" readership studies to prove that their figures are measuring actual readership.

Psychological Factors in Instrument Reading. II. The Accuracy of Pointer Position Interpolation as a Function of the Distance Between Scale Marks and Illumination *

Walter F. Grether

Aero Medical Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio

and

A. C. Williams, Jr.

University of Illinois

The reader of instruments is normally expected to obtain values of greater precision than the graduations placed upon the instrument scale. To accomplish this he must interpolate, that is, estimate the relative distance of the pointer from the two scale marks between which it falls and assign an appropriate value to this position. The accuracy with which this can be done obviously limits the precision with which any given scale can be read. The accuracy of such interpolation, moreover, will be influenced by several variables in the scale design and the conditions of reading. For a prediction of reading precision obtainable with different instrument designs under various conditions of viewing the effect of the significant variables must be known. In the present experiment the accuracy of pointer position interpolation was studied as a function of (a) diameter of the dial; (b) angular separation of the scale divisions; and (c) simulated day- versus night-viewing conditions. It will be shown in the presentation of the results that the first two variables can be reduced to a single one, namely, the length of the arc (in visual angle or inches) between scale marks.

A problem in scale design to which the present investigation is particularly relevant is concerned with the question of how finely a scale should be divided in order to provide maximum reading accuracy. In an investigation by Loucks (5) the legibility of tachometer dials was investi-

gated using rather short exposure (0.75 sec.). For three dials with graduations of 100, 50, and 20 RPM respectively the percentage of reading errors increased as the value and size of the graduations decreased. From this finding it might be concluded that placing the graduations rather close together will decrease rather than increase reading accuracy. The findings of Kappauf, Smith, and Bray (3), and Kappauf and Smith (4), in experiments where the exposure interval was not limited, disagree with those of Loucks. In their experiments dials graduated in units gave greater reading accuracy and speed than dials of the same size but graduated in 5- or 10-unit steps. However, the superiority of 1-unit over 5-unit graduations was rather small and not at all in proportion to the increased number of graduation marks. These latter results are in agreement with those of an investigation by Grether (1) on the reading of clock dials. With one minute as the criterion of reading accuracy, dials with 1-minute graduations gave higher reading accuracy than similar dials with only 5-minute scale marks.

A possible explanation can be offered for the discrepancy between the findings of Loucks (5) and later investigators. It is quite probable that as the number of scale marks is increased more eye fixations are required to make each reading. By limiting the exposure time and consequently the number of eye fixations Loucks may have favored those dials with more widely spaced graduations.

In the study of scale designs it is helpful to distinguish between two general types of errors encountered in dial reading studies. There are first the precision errors or errors of interpolation. These can never exceed in magnitude the value of the smallest interval on the scale. The other type may be called comprehension of interpretation errors, in which an incorrect value is assigned to the graduation mark against which the pointer is being read. Comprehension errors are frequently very large and are usually some multiple of the minor, intermediate, or major scale divisions. In a study by Grether (2) of altimeter reading, for example, most of the errors were of this latter sort, with errors of 1000 feet being particularly common. It is important to recognize that many of the dial reading studies up to the present have been concerned only with the interpolation type of errors, when in actuality the larger comprehension errors are far more serious in practical instrument reading situations. It is quite possible that the presence of a large number of graduation marks on a dial may greatly increase the probability of large comprehension errors and thereby nullify the precision which a finely graduated scale makes possible.

The spacing of the divisions on a scale is usually not a variable concerning which the instrument designer has a free choice. Normally the

physical length of the scale, the range of values to be covered, and the desired accuracy of reading are fixed by the particular application. Based upon these requirements the designer must then select values (usually 1, 2, 5, or decimal multiples of these) for his scale increments which will give him reasonable spacing between graduation marks.

The aim of the present study was to provide the instrument designer with data from which to predict how the accuracy of readings will be affected by the physical length of the interval into which he sub-divides a scale. Measurements were made under two lighting conditions comparable to those under which aircraft instruments are viewed. Emphasis in this study was placed on interpolation errors. The more complex comprehension type of errors were recorded but were relatively few in number and were not subjected to analysis.

Apparatus

For the purpose of this experiment a series of 16 simulated instrument dials was prepared. A sample dial and pointer are shown in Figure 1. Four sizes of dials were used as follows: 1, 1⅝, 2¾, and 4 inches in diameter. The particular dimensions of the two intermediate sizes were chosen to duplicate standard aircraft instruments. Each size of dial was produced with four different graduation intervals, defined in terms of the angular separation between scale marks, as follows: 5, 10, 20, and 40 degrees. Except for the variations in diameter and size of graduation intervals, all dials were identical. The intermediate graduation marks were ⅛ inch in length and approximately 0.02 inches in width. The major graduation marks at each end of the scale were the same width but ¼ inch in length. The numerals on all dials were ⅛ inch in height. All pointers were 3/32 inch in width and of such a length that the tip reached to the inner edge of the shortest graduation marks. All dials covered a range of from 0 to 50 units as shown in Figure 1, with graduation marks only at the 0, 10, 20, 30, 40, and 50 positions, and numerals only at 0 and 50. These dials were engraved on brass plates, which were then painted a flat black and the engraved markings filled with yellow fluorescing paint (pale yellow in daylight) as used on the latest type of USAF instruments.

The experimental dials were presented singly in a panel opening 30 inches from and perpendicular to the subject's eyes. Daylight conditions were simulated with a fluorescent type daylight lamp which provided an illumination of 45 foot-candles at the panel opening. For simulation of night conditions the subject's room was completely darkened and the dial illuminated with a standard C-5 ultra-violet aircraft instrument panel light operating at maximum intensity. No means were available for

obtaining a quantitative measurement of the brightness of the scale markings under ultra-violet illumination. Covering the opening in which the experimental dials were presented was a mechanical shutter operated by the experimenter.
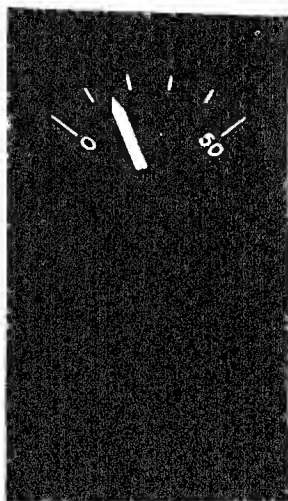
5 I 0 4 - E

FIG. 1. Sample dial and pointer (1⅞ inch diameter and 20 degree angular separation between scale marks).

On the experimenter's side of the test panel was a carriage on which four of the dials could be mounted side by side. This carriage rode upon two horizontal tracks parallel to the screen. To present any one of the dials the experimenter moved the carriage so that the desired dial would appear in the panel opening. At the experimenter's side of the carriage were four master setting dials 5 inches in diameter. On each of these dials was a pointer connected to the same shaft as the pointer on the dial to be read by the subject. On the experimenter's dials were closely spaced graduations which made possible accurate settings to one-tenth of the space between graduations on the subject's dials.

Also provided at the experimenter's station was a lever for manual operation of the shutter used to expose the dial to the subject. This lever was used also to operate an electric timer through a suitable switch. Thus, the timer indicated the time during which the shutter remained open. Since the experimenter closed the shutter as soon as the dial reading had been completed, the reading on the clock gave a crude measure of the reaction time on each test trial. Several other methods of measuring reaction time were tried but found to be unsatisfactory.

Eighty male college students were used as subjects in this experiment. Only men with 20-20 binocular vision (corrected or uncorrected) were accepted. The subjects were seated in a chair in front of the screen with their eyes 30 inches from the panel opening and with the line of sight perpendicular to the panel opening in order to eliminate parallax. The subjects were divided into groups of 20, each group being tested on a set of four dials. The four dials included one of each diameter and one of each graduation interval. Each subject was given a total of 80 trials, equally divided among the four dials in a random sequence. Of each group of 20 subjects, 10 were tested under simulated daylight conditions and the remaining 10 under simulated night conditions.

A variety of dial settings were chosen so as to represent all portions of the dial from 0 to 50. The actual numbers to be read were the same for all dials although the order of presentation was randomized. The subjects were instructed to read the dials as quickly and accurately as possible to the nearest whole number. As can be seen in Figure 1, the reading to the nearest whole number required estimation to the nearest one-tenth of the distance between graduations.

On each trial the experimenter set the pointer of the dial to be presented, then opened the shutter and waited for the subject's verbal response, following which the shutter was closed and the subject's reading and the clock score recorded.

Results

The experimental design resulted in 200 readings on each of the 16 specific dials under each of the lighting conditions. For each reading both error and time data were obtained. The error data consisted of the deviations of the readings from the actual settings. These deviations could be either negative or positive and increased in step intervals of one, or one-tenth of the space between graduation marks. For purposes of analysis, however, error distributions were made without regard to sign. Since the distributions of these errors, and also response times, were considerably skewed, with the modal error being zero for many of the dials, the statistical treatment presented in this report is limited to medians and 75th percentiles. Means were computed for all the data and found

to present the same general picture as the medians, but the values were inflated because of the skewness.

A summary of the error data for the 16 dials is shown in Table 1. In the third column, it will be noted, the two variables of dial diameter and angular spacing of the divisions have been reduced to a single variable, namely, the length of graduation interval defined as the arc between the inner ends of the shortest scale marks. This value can be described also as the distance the pointer tip must travel between adjacent graduations.

Table 1
Summary of Data on Accuracy of Pointer Interpolation as Function of Dial Diameter and Spacing of Scale Divisions

Dial Diameter, inches	Angular Spacing, degrees	Length of Inner Arc, inches*	Median Error Daylight, % of interval	Median Error Night, % of interval	Median Error Combined, % of interval	75th Percentile Error, Combined, % of interval	Median Error Combined, degrees
1	5	.032	21.8	31.0	26.4	45.5	1.32
1	10	.065	17.8	18.8	18.3	28.5	1.83
1	20	.130	14.3	14.9	14.6	21.2	2.92
1	40	.261	13.2	12.1	12.6	17.6	5.04
1⅝	5	.070	20.0	19.4	19.7	31.8	0.99
1⅝	10	.141	11.3	14.2	12.8	18.5	1.28
1⅝	20	.238	12.4	10.3	11.4	17.1	2.28
1⅝	40	.567	7.8	8.1	8.0	13.0	3.20
2¾	5	.109	15.4	15.7	15.5	21.2	0.78
2¾	10	.218	12.0	9.3	10.6	16.5	1.06
2¾	20	.436	9.4	9.1	9.3	15.2	1.86
2¾	40	.872	8.1	8.1	8.1	14.1	3.24
4	5	.163	14.9	13.9	14.4	20.0	0.72
4	10	.327	9.7	9.0	9.3	15.2	0.93
4	20	.654	8.8	7.9	8.3	14.3	1.66
4	40	1.309	9.8	9.1	9.4	15.0	3.76

* For conversion to minutes of visual angle multiply by 114.7.

This length of the inner arc is presented in inches with the multiplying factor provided for conversion to minutes of visual angle for the 30-inch viewing distance. A comparison of the columns of median errors for daylight and night conditions in Table 1 reveals no consistent difference between these two lighting conditions. Only for the dial with the most closely spaced divisions does the performance under daylight appear to be superior. For this reason it seemed safe to combine the two sets of error data in the remaining columns of the table.

Some of the most important findings contained in Table 1 are presented graphically in Figures 2 and 3.   With length of graduation interval along the base line, the median and 75th percentile errors are plotted as per cent of the interval in Figure 2, and as absolute values in Figure 3.
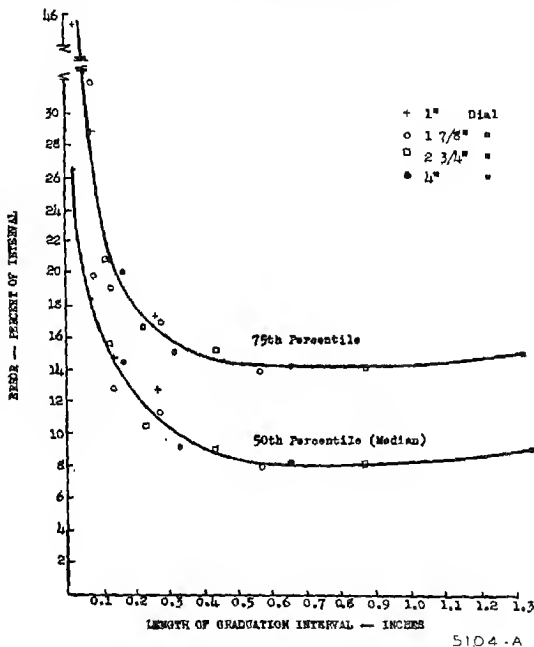

FIG. 2.   Relative error of interpolation as a function
of length of graduation interval.

Figure 2 will be recognized as a typical Weber function in which threshold ratios (DI/I) are plotted as a function of the stimulus intensity (I).   In this figure, it will be noted, the relative accuracy of interpolation is very nearly constant for graduation intervals above 0.5 inch, with a slight rise for the largest interval.
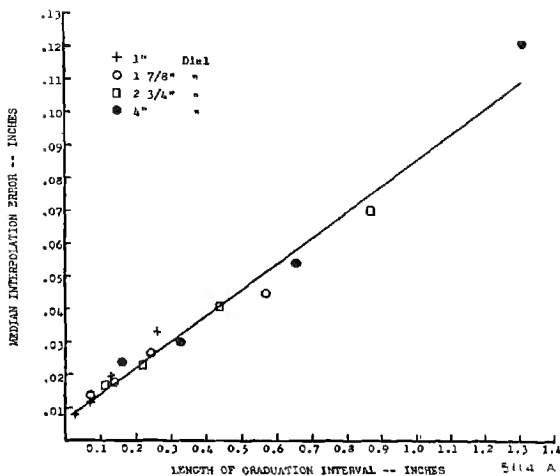
Fig. 3. Absolute error of interpolation as a function
of length of graduation interval.

Table 2

Median Time for Interpolation of Pointer Position as a Function
of Dial Diameter and Spacing of Scale Divisions

Dial diameter, inches	Graduation Interval			
	5°	10°	20°	40°
	Seconds per reading for daylight conditions			
1	1.98	1.78	1.91	1.84
1⅞	1.73	1.80	1.86	1.76
2¾	1.83	1.85	1.73	1.77
4	1.87	1.75	1.68	1.90
	Seconds per reading for night conditions			
1	2.48	2.26	2.10	2.13
1⅞	2.18	2.14	2.02	2.10
2¾	2.15	2.05	2.02	2.06
4	2.00	2.02	2.08	2.23

In Figure 3 the absolute values for interpolation thresholds fall very nearly on a straight line, which if extrapolated to zero intervals would intercept the ordinate at approximately 0.006 inch. It is apparent from this curve that no limit had been reached for absolute accuracy of interpolation in this experiment. The excellence with which the various points fit the curves in these two figures indicates that the combination of dial diameter and angular separation into the single variable of length of graduation interval was justified.

The results of the measurements of response time in this experiment are summarized in Table 2. It is apparent that there are no consistent relationships between response times and the dial dimensions, although the method of measuring response time may have been too crude to demonstrate minor relationships that might have been present. On the other hand the response times for the night viewing conditions are consistently higher than under the daylight conditions. It is quite possible that this latter finding was an artifact resulting from a slight delay between opening of the shutter and the fluorescing of the scale marks.

Discussion

Effect of Illumination on Interpolation Accuracy. It is apparent from Table 1 that there was no difference in accuracy of dial reading for day and night conditions except for the dial with the most closely spaced graduations. This finding is in general agreement with results obtained by Spragg and Rock (6) in an investigation of the accuracy of interpolation as a function of illumination. These investigators found accuracy to be almost constant down to a scale mark brightness of 0.022 foot-lambert. In the present experiment the brightness of the scale markings under the simulated night conditions is estimated to have been considerably above the 0.022 foot-lambert value below which Spragg and Rock found a marked loss in accuracy of interpolation.

Effect of Separation between Scale Marks on Speed of Reading. It is noteworthy that in this experiment no relationship was found between the space between graduation marks and speed of reading, although it is admitted that the method of measuring speed of reading lacked precision. In experiments conducted at Princeton University reading time increased as the space between marks was decreased, but in all cases there were changes in other variables which could have caused the changes in reading speed. In the first experiment by Kappauf, Smith, and Bray (3) the reduction in space between scale marks was accompanied by reductions in all other dial dimensions. In the second experiment by Kappauf and Smith (4) reduction of the space between marks was accompanied in some cases by reduction in all other dimensions, in other cases by an increase

in total range of values covered by the scale. The question of whether or not the separation between scale marks in isolation has any effect on speed of dial reading does not appear to have been answered definitely.

Effect of Dial Dimensions on Relative and Absolute Accuracy of Interpolation. In discussing the findings regarding accuracy of interpolation the distinction must be constantly kept in mind between accuracy relative to the interpolation space, and accuracy in absolute units such as degrees or inches. It is apparent from Fig 2 that there is scarcely any useful gain in relative accuracy of interpolation as the graduation intervals are increased beyond 0.25 inch. On the other hand as the intervals are reduced below this value relative accuracy falls off very rapidly. Approximately one-fourth (0.25) to one-half (0.50) inch would therefore seem to be an optimum value for graduation intervals from the standpoint of relative accuracy.

For maximum accuracy in an absolute sense it would appear from Figure 3 that the optimum graduation interval, if there is such, is below the range covered by the present experiment. The data in Figure 3 suggest that the absolute value of interpolation errors might continue to decrease with decreases in graduation interval until the limit of visual acuity·is reached. If this is true the limit of visual acuity would determine the optimum graduation interval for maximum accuracy of dial reading. The data of Loucks (5) and Kappauf, Smith, and Bray (3) suggest that as the distance between graduation marks is decreased, there is an increasing tendency to make comprehension errors, that is, assign the wrong values to scale marks. Also, Kappauf and Smith (4) have found that increasing the total number of marks on the scale increases the time required for reading. It would seem, therefore, that there is no easy answer to the problem of what is the optimum interval size for instrument scales, but that the optimum interval will vary with reading criteria.

Summary

Measurements were made of the accuracy of interpolating pointer position between scale marks as a function of dial diameter and the angular spacing between divisions. Subjects were required to estimate the pointer position to within one-tenth the space between graduations. The experimental dials were painted with yellow fluorescing paint on a black background and were read under simulated daylight (45 foot-candles) and night (ultra-violet) illumination conditions. The major results of this investigation may be summarized as follows:

1. Dial diameter and angular spacing of the scale marks could be combined into the single variable of length of graduation interval.

2. The relative error of interpolation decreased as the length of the

graduation interval increased up to approximately 0.5 inch, and was very nearly constant at higher intervals (see Figure 2).

3. The absolute error of interpolation increased very nearly as a linear function of the length of the graduation interval (see Figure 3). If there is an optimum interval for absolute accuracy it would appear to be below the interval lengths used in this study.

4. Except in the case of the most closely spaced divisions the accuracy of interpolation was independent of the two illumination conditions.

5. The speed of dial reading was not systematically related to either dial diameter or angular spacing of the divisions, although the measurements were admittedly crude. Slower reading under the simulated night (ultra-violet) lighting conditions was probably due in part to delay in fluorescence of the dial markings.

References

1. Grether, W. F. Factors in the design of clock dials which affect speed and accuracy of readings in the 2400-hour time system. *J. appl. Psychol.*, 1948, *32*, 159–169.

2. Grether, W. F. Psychological factors in instrument reading. I. The design of long-scale indicators for speed and accuracy of quantitative readings. *J. appl. Psychol.*, 1949, *33*, 363–372.

3. Kappauf, W. E., Smith, W. M., and Bray, C. W. *Design of instrument dials for maximum legibility. I. Development of methodology and some preliminary results.* USAF Air Materiel Command Memorandum Report No. TSEAA-694-1L, 20 October 1947.

4. Kappauf, W. E., and Smith, W. M. *Design of instrument dials for maximum legibility. II. A preliminary experiment on dial size and graduation.* USAF Air Materiel Command Memorandum Report No. MCREXD-694-1N, 12 July 1948.

5. Loucks, R. B. *Legibility of aircraft instrument dials: The relative legibility of tachometer dials.* AAF School of Aviation Medicine, Project No. 265, Report No. 1, 1944.

6. Spragg, S. D. S., and Rock, M. L. *Dial reading performance as related to illumination variables. I. Intensity.* USAF Air Materiel Command Memorandum Report No. MCREXD-694-21, 1 October 1948.

Identification of Cola Beverages. III. A Final Study *

N. H. Pronko and J. W. Bowles, Jr.

Wichita University

In a preliminary study, the authors (2) found that when four Cola beverages (Coca Cola, Pepsi Cola, Royal Crown Cola and Vess Cola) were presented to Ss to identify, the identification responses of those Ss tended to cluster around the three brand names, Coca Cola, Pepsi Cola and RC Cola, the fourth sample being consistently labelled with one of the three names previously employed. This occurred both when Ss were given four samples each of *different* Cola beverages or four samples of the *same* Cola.

It was therefore decided to employ only three different Colas on the hypothesis that since our Ss apparently did not discriminate the Colas on a gustatory basis, their identification responses would be distributed among the three brands in an order approximating chance. Such a procedure was carried out in Experiment II (1) which showed that whether Ss were given three different beverages or the same beverage three different times, their identifications were not essentially different in the two cases. It was concluded that when Ss were asked to identify the three leading brands of Cola, they might just as well have drawn their names out of a hat.

Another experiment suggested itself as the logical outcome. If Coca Cola, Pepsi Cola and RC Cola identification responses were randomly distributed when these beverages were actually given serially or otherwise, what would the distribution of those responses be when some relatively unknown Cola drinks were given *instead* of those three? Therefore, the present experiment utilized three brands of less well known or advertised Cola. These were Hyde Park Cola, Kroger Cola, and Spur Cola. The last beverage is now, for the first time, being distributed in this community although none is as generally available as the three leading brands.

Procedure

As in Experiment II, the present study employed two groups of Ss, 96 in Part I and 60 in Part II. For the most part, these were beginning students in Elementary Psychology.

Part I. Each of 96 Ss was admitted individually into the experimental room and was invited to sit down. The following instructions were then read to him.

We would like to have you taste and identify some Cola drinks. You will be told in what order and when you are to drink them. After you have finished each sample, report your identification to E and take enough water from the paper cup to rinse your mouth well.

A tray containing three one-oz. glasses of Hyde Park Cola, Kroger Cola and Spur Cola respectively were placed before the *S.* He was then told to drink the beverages labelled x, y, and z in the order indicated to him. Samplings were spaced about a minute apart, *S*'s name and other information being recorded in the interval between drinks.

Order of presentation of beverages, determined preexperimentally, was such that the three stimuli appeared in the first, second and third position 32 times. Such a counter-balanced order was used in order to preclude the operation of position effects or stimuli interactions orally. All beverages were at all times kept out of sight of Ss and were placed in a refrigerant maintained at approximately 5° C.

Part II. In part II 60 Ss were administered *the same* Cola drinks at each of three trials. Thus, 20 were given all Hyde Park Cola; 20, all Kroger Cola; and 20, all Spur Cola.

Results

Table 1 shows the distribution of the identification responses of the 96 Ss of part I who were given one-oz. samples each of Hyde Park Cola, Kroger Cola and Spur Cola. The most conspicuous finding to be observed here is the total absence of correct identifications! Not on a single occasion were any of these beverages correctly named. As a matter of fact, it will be noted that the greater proportion of identification responses are to be found in the Coca Cola, Pepsi Cola and RC Cola columns with a sprinkling of namings under Seven Up, Dr. Pepper and Cleo Cola.

Table 1

Showing the Distribution of 288 Identification Responses when Each of the 96 Ss was Presented in Turn, but in Counterbalanced Order, with One 1 oz. Sample Each of Hyde Park, Kroger, and Spur Cola

| Brand of Beverage Given S | Frequency of Ss' Identification Responses | | | | | | | | | | |
	H.P.	K.	S.	C.C.	Pep.	R.C.	7 up	Dr. Pep.	Cleo.	Other	Total
Hyde Park	0	0	0	39	27	24	1	1	0	4	96
Kroger	0	0	0	30	39	22	1	0	1	3	96
Spur	0	0	0	34	33	22	2	1	1	3	96
Total	0	0	0	103	99	68	4	2	2	10	288

Our *S*s in this experiment show essentially the same behaviors as those of Experiment II who were actually given Coca Cola, Pepsi Cola and RC Cola. Our hypothesis is, therefore, substantiated and we must conclude that our *S*s have not discriminated from among the varieties of Cola beverages employed in our series of studies. Instead, they have applied a readily available repertoire of naming reactions whose probable source is advertising, familiarity through actual contact and other forms of culturization. One more point should be noted. In the total of 288 identification responses, observe that the three relatively less well known beverages here employed were identified as Coca Cola 103 times, as Pepsi Cola 99, and as RC Cola 68 times. Again, as in previous studies, Coca Cola and Pepsi Cola names are applied with greatest frequency, with RC trailing as before.

Table 2

Showing the Distribution of 180 Identification Responses when Each of 60 *S*s was Presented with Three 1 oz. Glasses of Either the Single Brand, Hyde Park, Kroger, or Spur Cola

Brand of Beverage Given S	Frequency of *S*s' Identification Responses										
	H.P.	K.	S.	C.C.	Pep.	R.C.	7 up	Dr. Pep.	Cleo.	Other	Total
Hyde Park	0	0	0	25	21	12	0	1	1	0	60
Kroger	0	0	0	22	25	13	0	0	0	0	60
Spur	0	0	0	26	19	9	0	1	2	3	60
Total	0	0	0	73	65	34	0	2	3	3	180

Table 2 shows distribution of the 180 identifications of our 60 *S*s who were presented with three samples of the *same* Cola. The reader will note that whether our *S*s get three samples of Hyde Park, Kroger or Spur Cola, they nevertheless identify each of them most frequently as Coca Cola and Pepsi Cola and less frequently as RC Cola with the same sprinkling of other brands as before. The pattern of total frequencies is the same as in all parts of this and other experiments with Coca Cola and Pepsi Cola in the lead and RC as runner-up. We suggest that this consistent pattern may reflect the relative efficacy of the advertising of these three brands.

Summary

A group of 96 *S*s was asked to identify one-oz. samples of Hyde Park Cola, Kroger Cola and Spur Cola presented in counterbalanced order.

1. There were *no* correct identifications. Instead, these beverages were identified 103 times as Coca Cola, 99 times as Pepsi Cola, 68 times as RC and a total of 18 times as some other soft drink.

2. Another group of 60 *S*s, each of which was given three one-oz. samples of only one of the three Colas, showed an assortment of Coca Cola (73), Pepsi Cola (65) and RC Cola (34) identifications, again indicating total absence of correct judgments.

3. It is concluded that no matter whether our *S*s are asked to identify three or four of the *same* or *different* Cola beverages, regardless of what particular brands are employed, their identification responses are inevitably "Cola Cola, Pepsi Cola, RC Cola."

4. The seven brands of Cola beverages employed in our series of studies appear to have the same stimulus function for our *S*s and may be said to be "equivalent stimuli."

References

1. Bowles, J. W., Jr., and Pronko, N. H. Identification of Cola Beverages. II. A further study. *J. appl. Psychol.*, 1948, **32**, 559–564.

2. Pronko, N. H., and Bowles, J. W., Jr. Identification of Cola beverages. I. First study. *J. appl. Psychol.*, 1948, **32**, 304–312.

Book Reviews

Samuel A. Stouffer et al. *The American soldier: Volume I, Adjustment during army life; Volume II, Combat and its aftermath.* Princeton: Princeton University Press, 1949. Pp. 600 each. Vol. 1 and 2, $13.50. Separate, $7.50.

These are the first two of four volumes prepared and edited under the auspices of a Special Committee of the Social Science Research Council. The material reported is based upon data concerning soldiers' attitudes collected by the Research Branch of the Information and Education Division, War Department, during the period, December, 1941, to August, 1945.

The basic tool which the Research Branch used was the questionnaire survey. A typical survey went through the following stages: (1) a request for information concerning attitudes or behavior of soldiers by some agency or branch of the army; (2) conferences between members of the Research Branch and the requesting agency in which the area to be investigated was outlined; (3) a field trip by two or three members of the Research Branch in which casual conversations were held with enlisted men and officers on the topics to be covered in the survey; (4) preparation of a questionnaire based upon the experience gained in the field trip; (5) a pre-test of the questionnaire, followed by revisions; (6) final conferences with the requesting agency on the revised questionnaire; (7) clearance of the questionnaire with the Director of the Research Branch; (8) sending the questionnaire into the field.

Samples for domestic surveys were selected in Washington; those for overseas surveys were selected in theater headquarters. In general the questionnaires were filled out directly by the men, personal interviews being limited to those in the lower educational groups.

Over 200 such surveys were made by the Research Branch in which practices, preferences, and attitudes were investigated. The following partial list indicates something of the diversity of the areas surveyed: use of atabrine; practices related to trench foot; housing preferences of troops stationed in Alaska; popularity of various departments in *Yank;* radio listening habits; attitudes toward the British, Chinese, and Germans; leisure-time activities; attitudes toward the WACS, civilians, Negroes, and MP's; practices in relation to venereal disease prevention; attitudes toward medical care, hospitalization, war, savings and insurance plans, rotation and demobilization; reactions to army films, orientation and

indoctrination programs; attitudes toward service in the tropics, combat duty, job assignment, and training; reactions to German weapons; desires for educational courses; preferences in Christmas gifts; and studies of map reading, word pronunciation, and psychoneuroses.

Attitudes of special groups on various problems were surveyed. Some of the groups investigated were: Negroes, combat troops, Special Service officers, MP's, WACS, psychiatric patients, combat infantrymen, paratroopers, AWOL's, hospital nurses, combat veterans, Air Force returnees, combat veterans in hospitals, B-29 officers and enlisted men, and hospital patients.

As a result of these surveys, some concrete actions were taken by the army. One example: the point system for demobilization was based upon a survey of the factors which enlisted men thought should be considered and the relative weights which they thought should be assigned to these factors. In addition, a monthly bulletin, *What the Soldier Thinks*, was published and circulated down to regimental commanders. Another bulletin, the *Monthly Progress Report*, was prepared for higher level officers. Both bulletins were intended to keep officers informed concerning the attitudes of enlisted men and officers on various problems of interest to the army.

Although the first two volumes are primarily intended for the army, historians, social psychologists, and sociologists, they contain much useful information for the industrial psychologist and personnel workers. This is indicated by the following topics discussed in some detail in the various chapters: job satisfaction and job morale; leadership and social control; motivation; social mobility; adjustment; promotion; training. Most of these subjects are covered in Volume I which deals with the period of adjustment during army life. Volume II is primarily concerned with combat and its aftermath. Since many readers of this journal will be primarily interested in Volume I, it is unfortunate that separate indexes for the two volumes were not prepared. As it is, a single index is included in Volume II.

It is worth mentioning also that the index does not contain entries for the pertinent problems of *reliability* and *validity*, although some attempts at validation of the questionnaires were made and are discussed in the volumes. Another point of minor irritation is the frequent forward reference in Volumes I and II to the material to be published in Volumes III and IV.

Charts, graphs, and tables are to be found on almost every other page. Some of these are very elaborate and detailed. Others show the percentage of various groups responding to a single question in a particular manner. The basic statistic is the percentage, although a few correlation co-

efficients are reported and an occasional chi square is to be found—usually in a footnote. This is not intended as a criticism. The reader should keep in mind that the work of the Research Branch was primarily directed toward the collection of information about attitudes which would be useful to the army planners of information, orientation, and educational programs. Their job was, as one of the authors calls it, a "practical engineering" job. From the evidence reported in these two volumes, it was a job well done.

Allen L. Edwards

The University of Washington

Planty, Earl G., McCord, William S., and Efferson, Carlos A. *Training employees and managers for production and teamwork.* New York: Ronald Press, 1948, pp xiii + 278. $5.00.

This book is characterized by breadth of viewpoint, purpose, scope, and principles. Training is conceived as a training of the whole employee (on any job level) in attitudes, skills, and knowledges in such way that he fits into the framework of his job, his company, and the American system of free enterprise. The ambitious program undertaken by the authors could easily have led to a superficial or biased book. For the most part these potential dangers have been avoided. The authors indicate considerable evidence of sound professional training, practical experience, common sense, and penetrating insight. The result is a book which will be of considerable value and interest to all persons directly or indirectly engaged in training in business and industry.

The first two chapters cover what training is and what it will do. The second part, consisting of the next six chapters, deals with organizing, installing, and administering a training program. Basic principles are given together with specific techniques and mechanics. The third part of the book contains thirteen how-to-do-it chapters devoted to training programs and methods. Chapters are included on teaching aids, teaching in business and industry, the training staff, the small company, helpful resources, and training for specific groups such as new employees, supervisors, technical and professional, trade and semiskill, and office employees.

It was difficult for this reviewer to remain objective while appraising the book. The enthusiasm aroused by the first few chapters kept increasing throughout the entire book except for the chapter on technical and professional training. This chapter is of little value to the book or the reader except for the implication that such training has lagged far behind other types. A serious error of omission is the failure of the book to discuss training costs. Such a chapter would appear imperative in a book of this type.

The authors have shown good judgment in striking a balance between various topics, have organized the material excellently, and have used well chosen examples and illustrations. They are to be commended for attacking some of the cliches currently held by training men; as, for example, the universal superiority of conferences as compared with lectures. The choice of words and mechanics of writing are considerably better than usually found in books of this type. These factors all add up to the best book on industrial training which this reviewer has ever seen.

Clifford E. Jurgensen

Minneapolis Gas Company

deFord, Miriam Allen. *Psychologist unretired, the life pattern of Lillian Martin.* California: Stanford University Press, 1948. Pp. 130. $3.00.

Lillian Martin, a most versatile psychologist, came from a distinguished family, was self-supporting in her education, took her undergraduate work at Vassar, spent four years in graduate study of psychology in Germany in the 90's, held a professorship of psychology in Stanford University from which she retired for age at 65 in 1916. She then turned to private practice and hospital service as a clinical psychologist and worked enthusiastically and with notable success for 27 years until she died at the age of 91.

This book will be welcomed by lovers of biography as a model of literary style, by all clinical psychologists, and by all students of guidance in the art of aging.

Carl E. Seashore

Winter Park, Florida

Kaback, Goldie Ruth. *Vocational personalities: an application of the Rorschach Group Method.* New York, Columbia University: Bureau of Publications, 1946. Teachers College Contributions to Education No. 924. Pp. xii + 116. $2.10.

This study tests the hypothesis that vocational choice is in part a function of personality, the differences in personality responsible for a given occupational choice being measurable by the Rorschach Group Method.

Group Rorschachs were obtained from 75 accountants, 75 pharmacists, 75 accountancy students and 75 pharmacy students. Brief life data sheets were filled in by the subjects. Mean ages were respectively 37, 37, 18, and 20 years, ranges being greater for the pharmacist groups than for their comparison accountant groups. Accountants averaged more education than pharmacists; students were similar to each other. Pharmacists had 16, accountants 13 years of work experience in their field. A miscellaneous vocational group numbering 108 was obtained from (ap-

parently) a co-worker and responses were compared with the four criterion groups studied. Reliability of Rorschach scoring was not reported.

Kaback finds that she can differentiate accountants from pharmacists by multiple correlation of .54 between the criterion and 24 Rorschach components. This r is too low for effective prediction. Accountants were significantly higher than pharmacists in W, d, R, P, O, Fc, M, FM, Fm, H+A, Aobj., Pl and Obj. Pharmacists were significantly higher on Fk and At. This means (roughly) that accountants were more productive (this factor alone seems to account for many of the differences shown) higher intellectually, more "socially sensitive—or aware of a need for social contact," more cautious, creative and original and perhaps more widely diversified in interests than pharmacists, while the latter showed more "anxiety." These differences, as the author points out, while statistically significant, are not of great practical value in vocational counseling.

For students, a multiple r against the criterion of .653 was shown (a respectable r, but not high enough for good prediction.) Accountants in training were higher than pharmacists in training on R, P, H+A, W, M, Fm, F, FC, Obj., D, Emb., and Arch; hence differences resemble the professional differences.

The author concludes: all groups normal, reasonably well adjusted (accountancy students best), accountants more intelligent, pharmacists less mature and impulsive (for both students and professional persons); and students of both groups show selves as more practical than the professional groups.

The author is to be commended for a type of study greatly needed. It seems however, first, that such a study can hardly attain highly reliable or valid results as long as socioeconomic pressures, family predilections, etc., determine the chosen career, often quite without regard to the personal wishes or capabilities of the individual involved; and that, second, a Rorschach study is difficult if not impossible when scoring categories of the test are considered (as is almost necessary) without the benefit of refinements in statistics of percentages, ratios, and sub-relationships.

The most ardent proponent of the Rorschach would probably not regard it as infallible when adapted as a group method and scored for means, using isolated scoring categories. In the hands of a competent technician, used to illuminate other data, it often seems to be of great use in vocational counseling; more "clinical" validations along the rather promising lines shown by the author may result from this interesting, carefully done, but practically not too useful study, the final conclusion of which is necessarily that there are all types of persons who enter or study in the fields of accountancy and pharmacy.

Boyd McCandless

The Ohio State University

Eysenck, H. J.   *Dimensions of personality.*   Cambridge, at the University
Press; New York; The Macmillan Company, 1947.   Pp. 308. $5.00.

This book is the result of a cooperative effort to discover the main
dimensions of personality and to define them operationally by means of
strict experimental, quantitative procedures.   About forty distinct re-
searches were carried out on some 10,000 normal and neurotic subjects by
a research team of psychologists and psychiatrists at the Mill Hill
Emergency Hospital, a war time neurosis center in London.   Dr. Eysenck
and his collaborators have turned to fruitful purpose the opportunities
such a place affords for studying some of the main factors of personality
and have shown how profitable may be the collaboration of psychologists
and psychiatrists in pursuit of this aim.

The first chapter entitled, *Methods and Definitions* describes the
methodological conditions underlying the researches, and the working
concepts and theories of temperament and personality structure.   The
second chapter, entitled, *Assessments and Ratings* presents a factorial
analysis of 39 trait ratings, carried out by psychiatrists on 700 neurotic
patients; also described is a study with questionnaires dealing with
"neuroticism," "persistence," and "irritability."   *Physique and Constitu-*
tion concerns studies performed on normal and neurotic subjects through
the methods of factorial analysis wherein main dimensions of body con-
figuration were isolated objectively; certain personality differences were
also found with respect to body size.   The fourth chapter, *Ability and*
Efficiency presents discussions and data on intelligence and neurosis,
"scatter," level of aspiration, perseveration, and persistence.   *Suggestibil-*
ity and Hypnosis attempts to distinguish various types of suggestibility,
to establish the relation between suggestibility and hypnosis, and to dis-
cover personality correlates of suggestibility (its relation to hysteria and
neuroticism).   Chapter six, *Appreciation and Expression* offers experi-
mental evidence on preference judgments.   *Synthesis and Conclusions,*
chapter seven, summarizes the researches indicating two main personality
dimensions namely, "neuroticism" and "extraversion-introversion,"
wherein the former is a general factor in the conative sphere, while the
latter is a general factor in the affective sphere.

While this book may presently contribute little that is useful to
harrassed personnel psychologists, these specialists should nevertheless
be familiar with the methods of investigation utilized.   Much that is
presented may influence our concepts of psychopathology in the future.
However, it is recommended to all applied psychologists working in
clinical fields and is "must" reading for the small corps of experimental
clinical psychologists.

                                                      Arthur Weider

University of Louisville School of Medicine,
     Dept. of Psychiatry, Div. of Psychosomatic Medicine

New Books, Monographs, and Pamphlets

The Harvard list of books in psychology. Gordon W. Allport et al. Cambridge: Harvard University Press, 1949. Pp. 77. $1.00.

Differential psychology. Revised edition. Anne Anastasi and John P. Foley, Jr. New York: The Macmillan Co., 1949. Pp. 894. $5.50.

Twilight in India. Gervee Baronte. New York: Philosophical Library, 1949. Pp. 382. $3.75.

Orientation et selection professionnelles par l'examen psychologique du caractere. Fr. Baumgarten. Paris, France: Dunod, Editeur, 92, Rue Bonaparte, 1949. Pp. 184. 680 fr.

Social psychology of modern life. Revised edition. Steuart H. Britt. New York: Rinehart and Co., Inc., 1949. Pp. 703. $4.50.

Labor relations and hindrances to full production. S. L. Burk. San Francisco: California Personnel Management Association, 1948. Pp. 18. $1.00.

The neurosis of man. Trigant Burrow. New York: Harcourt, Brace, 1949. Pp. 428. $7.50.

Reading manual and workbook. Homer L. J. Carter and Dorothy J. McGinnis. New York: Prentice-Hall, Inc., 1949. Pp. 120. $1.75.

Human relations in public administration. Alfred De Grazia. Chicago: Public Administration Service, 1949. Pp. 52. $1.50.

Criteria for the life history. John Dollard. Reprint. New York: Peter Smith, Publisher, 1949. Pp. 294. $3.25.

The supervisor's management guide. M. Joseph Dooher and Vivienne Marquis, Editors. New York: American Management Association, 1949. Pp. 190. $3.50.

We human chemicals. Thomas Dreier. Scarsdale: Updegraff Press, Ltd., 1949. Pp. 122. $2.00.

Mental testing. Florence L. Goodenough. New York: Rinehart and Co., Inc., 1949. Pp. 609. $5.00.

Readings in social security. William Haber and Wilbur Cohen. New York: Prentice-Hall, Inc., 1948. Pp. 634. $5.75.

Psychosexual development in health and disease. Paul H. Hoch and Joseph Zubin, Editors. New York: Grune and Stratton, 1949. Pp. 283. $4.50.

Group guidance, principles, techniques, and evaluation. Robert Hoppock. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 393. $3.75.

Motor performance and growth. A developmental study of static dynamometric growth. Harold E. Jones. Berkeley: University of California Press, 1949. Pp. 181. $3.00, cloth, $2.00, paper.

Helping students find employment. Forrest H. Kirkpatrick et al. Washington, D. C.: American Council on Education Studies, 1949. Pp. 37. $.75.

Studies in human behavior. Merle Lawrence. Princeton: Princeton University Press, 1949. Pp. 184. $3.50.

Explorations in personal adjustment—a workbook. George F. J. Lehner. New York: Prentice-Hall, Inc., 1949. Pp. 144. $1.50.

Mental hygiene in public health. Paul V. Lemkau. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 396. $4.50.

The psychology of personal adjustment. Second edition. Fred McKinney. New York: Wiley and Sons, Inc., 1949. Pp. 752. $6.00.

Selling performance and contentment in relation to school background. Albert C. Mossin. New York: Bureau of Publications, Teachers College, Columbia University, 1949. Pp. 166. $2.75.

The pre-election polls of 1948. Frederick Mosteller et al. New York: Social Science Research Council, 1949. Pp. 396. Paper, $2.50; Cloth, $3.00.

The nature-nurture controversy. Nicholas Pastore. New York: Columbia University Press, 1949. Pp. 213. $3.25.

Perception of symbol orientation and early reading success. Muriel Catherine Potter. New York: Bureau of Publications, Teachers College, Columbia University, 1949. Pp. 69. $2.10.

Human relations in an expanding company. F. L. W. Richardson and Charles R. Walker. New Haven: Yale University Labor and Management Center, 1948. Pp. 95. $1.50.

Letters to my son. Dagobert D. Runes. New York: The Philosophical Library, 1949. Pp. 92. $2.75.

Happiness for husbands and wives. Harold Shryock. Washington, D. C.: Review and Herald Publishing Association, 1949. Pp. 256. $2.50.

Rehabilitation of the handicapped. William H. Soden, Editor. New York: Ronald Press Co., 1949. Pp. 399. $5.00.

Personnel management for supervisors. Claude E. Thompson. New York: Prentice-Hall, Inc., 1949. Pp. 208. $2.95.

Selected writings from a connectionist's psychology. Edward L. Thorndike. New York: Appleton-Century-Crofts, Inc., 1949. Pp. 370. $3.50.

Working with people. Auren Uris and Betty Shapin. New York: The Macmillan Co., 1949. Pp. 314. Probable price, $3.50.

Experimental foundations of general psychology. Third edition. Willard L. Valentine and Delos D. Wickens. New York: Rhinehart and Co., Inc., 1949. Pp. 472. $3.00.

Community under stress. Elizabeth Head Vaughan. Princeton: Princeton University Press, 1949. Pp. 160. $2.50.

Personnel selection in the British forces. Philip E. Vernon and John B. Parry. London: University of London Press Ltd., 1949. Pp. 324. 20/-net.

Children with mental and physical handicaps. J. E. Wallace Wallin. New York: Prentice-Hall Inc., 1949. Pp. 549. $5.00.

Out-of-school vocational guidance. Roswell Ward. New York: Harper and Brothers, 1949. Pp. 155. $2.50.

Constructing classroom examinations. Ellis Weitzman and Walter J. McNamara. Chicago: Science Research Associates, 1949. Pp. 153. $3.00.

Supervisory training—why, what, how. ILIR Publications, Series A, Vol. 3, No. 3, Urbana: Institute of Labor and Industrial Relations, University of Illinois, 1949. Pp. 24. Gratis.

Recruitment, selection, and indoctrination of female clerical employees. Personnel and Training Series Research Project Report No. 4. New York: Scott Foresman and Co., Miss Edith Harper, 1949. Pp. 35. $2.00.

Sales manager's handbook. Chicago: The Dartnell Corporation, 1949. Pp. 1150. $10.00.

Training for tomorrow: proceedings of the Third Annual Training Conference of Educational Directors in Industry and Commerce. Montreal: Canadian Industrial Trainers' Association, 1949. Pp. 123. $2.00.

Executive personality and job success. New York: American Management Association, 1948. Pp. 35. $.75.

Building worker interest in production problems. New York: American Management Association, 1949. Pp. 34. $.75.

Management's role in industrial mobilization. New York: American Management Association, 1948. Pp. 27. $.50.

Organization controls and executive compensation. New York: American Management Association, 1948. Pp. 54. $1.00.

Personnel functions and the line organization. New York: American Management Association, 1948. Pp. 31. $.75.

Worker morale and productivity. New York: American Management Association, 1948. Pp. 38. $.75.